Does Implicit Bias Threaten the Semicompatibilist Position on Free Will and Responsibility?

David Gregory MA

Supervisors: Professor David Cockburn and Doctor Tristan Nash

Submitted in partial fulfilment for the award of the degree of Doctor of Philosophy

University of Wales Trinity Saint David

2021

DECLARATION SHEET

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ....David Gregory........................... (student)

Date ....20th April 2021............................

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used the extent and nature of the correction is clearly marked in a footnote(s). Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ....David Gregory.............................. (student)

Date ....20th April 2021..

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ....David Gregory............................. (student)

Date ....20th April 2021...

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for deposit in the University's digital repository.

Signed ....David Gregory.............................. (student)

Date ....20th April 2021..

Abstract

The nature and limitations of human freedom have been discussed for millennia and continue to be widely and enthusiastically debated today. This is not surprising, as our sense of freedom is felt to be an essential part of what it is to be human and many fundamental issues, such as responsibility, praise and blame are commonly considered to depend on how freely behaviour is undertaken. Understanding human freedom is valuable in itself and in terms of implications for other important areas, for example, the Law and Government policy.

This Thesis critically examines the semicompatibilist free will position in the context of implicit bias. Specifically, I answer the question, does implicit bias threaten the semicompatibilist position on free will and responsibility? A threat arises if behavioural expression of implicit bias is not subject to semicompatibilist conditions of agent control and so responsibility (guidance control) in the presence of compelling argument and substantial evidence supporting the counter position, that agents are responsible for such behaviour.

Responding to this question, I provide in Part I a brief historical perspective of the discussion of human freedom, followed by description of some major positions within the free will debate, focusing on semicompatibilism. Part II explores implicit bias in terms of what it is, how it is measured and implications for responsibility and control of influenced behaviour.

Having gained insight into semicompatibilism and established a model of implicit bias, Part III examines the impact of implicit bias on the semicompatibilist position, assessing and reaching conclusions concerning the ability of semicompatibilists to accommodate the phenomenon of implicit bias within their explanatory model. I then consider the implications of implicit bias for a particular defence of semicompatibilism from one of its major threats, the problem of moral luck.

I show that semicompatibilism successfully accommodates the phenomenon of implicit bias; agent responsibility for issuing behaviour is confirmed, in harmony with the presented models of implicit bias. A particular understanding of implicit bias is found to cause a problem for defence of semicompatibilism from the luck problem.

In response to the question, does implicit bias threaten the semicompatibilist position on free will and responsibility? I conclude that semicompatibilism, as a position on free will and responsibility, is immune from threats that originate from implicit bias.

Acknowledgements

I wish to extend sincere thanks and gratitude to my Supervisors, Professor David Cockburn and Doctor Tristan Nash, for invaluable guidance and support while writing this Thesis and during all my studies with The University of Wales Trinity Saint David.

I thank my wife Barbara, whose patience and support made this Thesis possible, and my parents, who would be very proud and delighted.

I am indebted to Bryan Magee (1930-2019) whose philosophical writing and broadcasting are a constant source of inspiration.

Contents

Part I – Free Will

Chapter 1

What is the Problem?

Chapter 2

Contemporary Responses to the Free Will Problem

Chapter 3

Semicompatibilism

Part II − Implicit Bias

Chapter 4

The Origin and Meaning of Implicit Bias

Chapter 5

Implicit Bias and Control

Part III − Semicompatibilism and Implicit Bias

Chapter 6

Does Implicit Bias Threaten the Semicompatibilist position?

List of Figures

Introduction

But Mathieu went on firing. He fired. He was cleansed. He was all-powerful. He was free.

Jean-Paul Sartre[1]

Our position on questions concerning the existence and nature of free will and responsibility has important theoretical and practical implications. Practical consequences concerning social policy, principles of justice and the basis of moral and judicial responsibility often originate from our understanding of human freedom. It is not an exaggeration to say, such practical consequences manifest every day, sometimes in life-or-death decisions concerning the development and implementation of policing and judicial sentencing policy.

There is much to discuss concerning what it means to act freely and the nature of free will and responsibility. It is a common assumption that a necessary condition for behaviour to be considered just, unjust, virtuous or corrupt is free choice; for an agent to act, for example with courage, the agent must be *responsible* for their actions and such responsibility entails free choice to act otherwise.[2] It will be seen such apparently uncontroversial thinking is far from universally accepted. Does the relatively recent phenomenon of implicit bias in certain situations reduce or remove completely moral responsibility for certain forms of behaviour by taking away or interfering in responsibility lessening ways with free choice? Under the influence of implicit bias, it is

[1] *Iron in the Soul* (Sartre 1963: 225). *Iron in the Soul* was published in 1949 (La Mort dans l'âme), the third volume of Jean-Paul Sartre's series of novels *The Roads to Freedom*. A fourth volume was published in 2011 *The Last Chance, Roads of Freedom IV*. (For the fourth volume, the translation of the series title *Les Chemins de la Liberté* was revised to *Roads of Freedom* rather than *The Roads to Freedom*; there is an obvious and important difference between freedom as a destination and freedom as an ongoing experience).

[2] A comprehensive survey of non-philosophers *For Whom Does Determinism Undermine Moral Responsibility? Surveying the Conditions for Free Will Across Cultures* (Hannikainen et al. 2019), spanning twenty countries and sixteen languages, shows interesting variations in common assumptions about necessary conditions, such as free choice and the possibility to act otherwise, for behaviour to be considered just or unjust, virtuous and responsible.

often claimed, an agent's actions in ethically relevant circumstances are influenced by factors that operate below the radar of consciousness. Typically, implicit bias is described as '… attitudes or stereotypes that affect our understanding, actions, and decisions in an unconscious manner' (Staats 2013: 6). It is claimed that if an agent is *unaware* of influencing bias and prejudice then associated actions are not freely and so not responsibly chosen because an agent is not *consciously*[3] aware of all relevant factors affecting their decision-making. Here, the form of interference with free choice (in responsibility lessoning ways) is claimed to be lack of awareness of motivating factors behind behaviour. If actions are not freely chosen, assigning responsibility, praise or blame for such actions appears to be wrong.[4] [5]Although not accepted by all, the most common characterisation of implicit bias is an agent displaying automatic and unconscious implicit associations that influence decision-making, often in negative ways, *particularly* judgment and evaluation of existing stereotypes or stigmatized groups. Implicit attitudes, 'likings' or 'dislikings', can be directed towards many things, including consumer products, but it is the 'very morally weighty' judgment and evaluation of existing stereotypes or stigmatized groups of *people*, and such discriminatory behaviour generally, which makes implicit bias *matter* (following Brownstein 2016a: 765). It is this surprising, perhaps alarming, influence of unconscious forces on decision making that appears so threatening to our understanding of free choice and responsibility for a considerable number of important, emotive and morally relevant actions: Is it possible

[3] Eddy Hahmias (2008) summarises widely, but certainly not universally agreed, conditions for free will:
(CR) Conscious Reflection: Agents have free will only if they have the capacity for conscious deliberation and intention-formation and that capacity has some influence on their actions.
(MR) Motivation by (potentially) endorsed reasons: Agents' free will is diminished to the extent their actions are motivated by factors that they are both unaware of and would reject were they to consciously consider them. Such conditions clearly have great resonance with issues relating to implicit bias and will be discussed shortly and throughout this Thesis.

[4] The relationship between concepts of freedom and responsibility is far from straight forward and is a central concern of this Thesis.

[5] There are issues here that need unpacking and clarification, for example, the claim that acting responsibly entails acting freely, requiring the possibility to act otherwise. Also, the idea that if an agent is unaware of influencing bias and prejudice then issuing actions are thereby not freely chosen because the agent is not consciously aware of all relevant factors, and so is not responsible. For an in-depth exploration of these issue see *Moral Responsibility and Consciousness* (King and Carruthers 2012), also *Consciousness, Free Will, and Moral Responsibility* (Caruso 2016).

to be responsible for actions influenced by factors that I am unaware of, have no control over and are therefore not freely chosen?

An essential issue is agent *awareness* of implicit bias and issuing actions. This is a complex and controversial matter; vitally, it will be found there are compelling arguments and substantial empirical evidence supporting the claim that agents *have* awareness of and *are* responsible for behavioural expression of implicit bias (BEIB)[6], *contrary* to the widely held view outlined above.[7] To produce a meaningful and valid critique of semicompatibilism within the context of implicit bias, models of implicit bias[8] will be described and chosen that incorporate *degrees* of agent awareness and responsibility for behaviour that is an expression of implicit bias. I defend this decision on the basis that a model that excludes *all* awareness and responsibility predicated on completely unconscious and automatic behaviour is essentially too simple *and* conflicts with considerable and substantial evidence and argument. This is recognised, I believe, by Brownstein and Saul when they describe implicit bias as ' … evaluations of social groups … *largely* outside conscious awareness or control' (added emphasis 2016a: 1). From the first shock of descriptions of implicit bias and related behaviour 'operating under the radar' of consciousness with complete absence of awareness and responsibility a more nuanced and defensible model of implicit bias is developed in Part II and taken forward into Part III where it is used to critically examine semicompatibilism. This describes in broadest strokes the landscape of this Thesis.

[6] I am grateful to Michael Brownstein for the term 'behavio(u)ral expression of implicit bias' (BEIB) (Brownstein 2016a: 768).

[7] Responsibility for behaviour issuing from implicit bias (BEIB) will be the dominant theme, but there are two further areas where responsibility and blame are relevant; responsibility for *having* implicit-associations and responsibility for responding appropriately to the knowledge that implicit-associations are part of our cognitive make up (following Holroyd 2012: 278).

[8] Please note well: The expression 'model of implicit bias' is used throughout this Thesis. Unless context clearly shows otherwise, this expression is used to describe a model of the mechanisms that mediate the influence of implicit social cognition, (i.e., attitudes, stereotypes), and *ex*plicit social cognition on behaviour, grouped in terms of impulsive and reflective elements; it is a model that shows behavioural expression of implicit bias mediated by various elements including the possibility of reflective decision making (Fig. 5.1).

Specifically, my main research aim is to explore and critically examine implicit bias and the free will position known as semicompatibilism[9] to answer the question; does implicit bias threaten the semicompatibilist position on free will and responsibility? I also consider the implications of implicit bias for a particular defence of semicompatibilism from one of its major threats, the problem of moral luck.

Responding to this objective, I present a brief overview of some major positions within the free will debate, focusing on semicompatibilism. I explore implicit bias in terms of what it is and how it is measured. Having gained insight into semicompatibilism and implicit bias, I examine the impact of implicit bias on the semicompatibilist position, assessing and reaching conclusions concerning the ability of semicompatibilists to accommodate the phenomenon of implicit bias within their explanatory model, and highlight any issues and areas for future research.

Part I briefly introduces a diversity of issues relating to human freedom and continues with increasing focus to examine contemporary responses to the free will problem. Literature relating to free will is vast and continues to increase rapidly, however, I have tried within Part I, and elsewhere, to resist many attractive diversions into interesting but peripheral areas. In Chapter 1 I reflect on a variety of free will related issues based on Ilham Dilman's (1999) historical survey and consider why free will has always been an important issue. Chapter 2 examines the main contemporary responses to the free will problem and Chapter 3 discusses the semicompatibilist position in greater detail. Part I provides an outline of free will and semicompatibilism sufficient to take forward into Part III.

The aim of Part II is similar to Part I; to achieve a clear understanding of implicit bias in preparation for Part III. The history and associated scholarship of implicit bias is large but considerably less than free will. While it is said that most well-known philosophers have had something to say about free will, this is not the case with implicit bias. Chapters 4 and 5 consider the origin and current meaning of implicit bias, and implicit bias and control. Implicit bias and the Implicit Association Test are described, leading naturally to description of Dual Process and Dual System theories of cognition.

[9] Following John Martin Fischer, for example, *Four Views on Free Will* (2007), I use semicompatibilism, rather than semi-compatibilism or Semi-Compatibilism.

Several approaches to implicit bias and responsibility are considered with view to finalising a position. A unified Dual System model is chosen based on the work of Deutsch and Strack (2010) together with the contrasting approach of Holroyd and Kelly (2016). Importantly, both approaches support the position that individuals are responsible for behaviour that has implicit bias as its source.

Having made necessary preparations within Part I and II, the main objective of the Thesis is addressed within Part III; to explore and critically examine the phenomenon of implicit bias as a threat to our contemporary understanding of free will and responsibility from the semicompatibilist perspective and consider the implications of implicit bias on a particular defence of semicompatibilism from the 'luck problem'. I believe this straightforward three-part approach is the clearest way to present the investigation and answers to these questions.

Appendix A describes the nature of mental representations responsible for biased behaviour, describing a position that is not built upon traditional propositional attitudes or associations. Appendix B outlines some aspects of agency and agent causation, as these issues are deeply integrated within discussion of free will and responsibility. Appendix C connects particularly with Chapter 1, providing a brief outline of Sophocles' *Oedipus Rex*. This Work illustrates ways in which human freedom may be restricted or eliminated by factors beyond an agent's control, yet the possibility of meaningful and responsible choices remains. This essential idea is present within semicompatibilism; the idea of guidance control, where the ability to do otherwise is unnecessary (regulative control), yet the possibility of making responsible choices is maintained.

I believe the important question, does implicit bias threaten the semicompatibilist position on free will and responsibility? has not been addressed previously. It is an interesting and important question because semicompatibilism is a major position within current mainstream compatibilism and implicit bias appears threatening to most plausible ideas of control and so responsibility because the influence of implicit bias on behaviour is, to a greater or lesser extent, outside conscious awareness. This is particularly problematic because our strong intuition is that typical resulting behaviour of implicit bias is something we *should* be taking responsibility for.

Using the three-part approach described, I argue and conclude that John Martin Fischer's semicompatibilism is not dangerously threatened by implicit bias: I confirm that implicit bias related behaviour *is* subject to guidance control and is therefore responsible behaviour, a conclusion *in harmony* with the presented model of implicit bias that is supported by significant theory and practice.

This is a vital point: There is substantial evidence supporting the claim that agents *are* responsible for implicit bias related behaviour. I investigate and show that semicompatibilism successfully accommodates the phenomenon of implicit bias. When challenged with implicit bias related behaviour the semicompatibilist model of free will and responsibility *endorses* agent responsibility for such behaviour. If this were not the case, some aspect(s) of semicompatibilism would require investigation and possible change. Endorsement of agent responsibility is confirmed, but a particular understanding of implicit bias was found to cause a problem for a defence of semicompatibilism from the luck problem. The essential contribution and conclusion of this Thesis: the semicompatibilist position concerning free will and responsibility is not threatened by the phenomenon of implicit bias.

Part I

Free Will

Chapter 1

What is the Problem?

Free will is arguably the most difficult problem in philosophy.
Susan Wolf[10]

The task is to formulate a conception of human action that leaves agents valuable; but what is the problem?
Robert Nozick[11]

1.0 Introduction

Chapter 1 presents reflections on free will and considers why free will has always been an important issue. I intend to introduce the free will debate, to build an historical viewpoint showing semicompatibilism as part of a long tradition of thought concerning the freedom and control human beings have throughout their lives.

1.1 Reflections on Free Will

For millennia there has been debate about the possibility and nature of human freedom. The debate continues, with many new articles appearing in scholarly journals and books every month. Much contemporary and recent historical discussion concerns the seeming incompatibility between the truth of physical determinism and human freedom, and the relationship between free will and moral responsibility. Adopting many forms and perspectives over the centuries, reflection on the nature of human freedom began over two thousand years ago within the works of some of the most important writers and

[10] *Freedom Within Reason* (Wolf 1993: vii).

[11] *Philosophical Explanations* (Nozick 1982: 291).

philosophers in western culture.[12] While perspectives change there is a constant theme; 'the extent to which human agents are in control of their own lives and destinies' (Russell 2013b: 2).

This section begins with some perspectives on the nature of human freedom, generally following Ilham Dilman's (1999) choice of writers and philosophers:

- Early Greeks: Sophocles (*Oedipus Rex* c429 BCE)[13] and Plato (*Gorgias* c380 BCE).
- Theological related issues within the works of St. Augustine (354-430 CE) and St. Aquinas (1225-1274).
- Free Will within Descartes (1596-1650), Hume (1711-1776) and Kant (1724-1804).
- Human freedom in the context of psychology with reference to works by Freud (1856-1939) and Sartre (1905-1980).

As mentioned, it is a sweeping yet plausible claim that every major philosopher has said something about free will, therefore my choice is to some extent arbitrary. I intend the work of the philosophers and writers described within this brief outline to represent some of the main perspectives within the historic free will debate. My purpose is not to give a detailed and comprehensive account, critique or comparative analysis but show briefly how the debate has assumed different forms and place later discussion in a wider context. Paul Russell, discussing Bernard Williams' *Shame and Necessity* (2008)[14] refers to the importance of historical context:

> Williams argues that methodologically, philosophers need to be historically sensitive and informed, and that typically they aren't. That's a problem with many discussions about free will, there is a lack of historical self-consciousness. (Russell 2013a)

[12] This Thesis only considers literature from within what is generally described as 'western culture'.

[13] See Appendix C for summary of Oedipus Rex.

[14] Bernard Williams' book *Shame and Necessity* (2008) will be mentioned again as it contains, as one would expect, interesting and insightful discussion of free will.

While providing 'historical self-consciousness' is greatly overstating my aim, it is important, I believe, to provide at least some background from which the contemporary debate can be seen to emerge and give context to Part II and III.

o Early Greeks: Sophocles, *Oedipus Rex*

The extent to which human agents are in control of their lives and destinies is examined in timeless epic form in Sophocles' *Oedipus Rex* c429 BCE. The idea of necessity is considered with the forewarning of events to Oedipus by the Oracle of Delphi.[15] The Oracle announced that Oedipus would incestuously father children with his mother and kill his father. As Dilman points out, it is interesting that while fulfilment of such a prophecy is clearly to be avoided, it is Oedipus' own actions while trying to avoid such a future that actually produce the events foretold (1999: 11). It is Oedipus who performs actions that ultimately lead to the prophecy being realised and so in this sense he has individual responsibility for those actions. The oracle's prophecy describes a future situation without detailing how it will be brought about. It is Oedipus' chosen actions that lead to the shocking outcome, actions that reflect Oedipus' character and so are, according to some philosophers, responsible actions. At the centre of this tragedy is a clash between inevitability of outcome and freedom of action. Without freedom of action the story loses all poignancy, but how can freedom truly exist in the fullest sense when a particular future state is inevitable? This is an important question and will be raised again shortly in the context of God's foreknowledge and free will. Dilman (1999: 14) makes the reasonable point that human freedom cannot exist in a total or absolute sense, that physical and logical constraints always apply, allowing action to take place only within certain boundaries that define what is within human control.[16] It is with a background of accepting such boundaries that freedom may legitimately be claimed to exist. Meaningful freedom is possible by focusing attention not on the impossible but on what can be done. Oedipus has freedom of this form, where constraints are those

[15] Delphi was an important ancient Greek religious sanctuary sacred to the god Apollo. Located on Mount Parnassus near the Gulf of Corinth, the sanctuary was home to the oracle of Apollo that gave cryptic predictions and guidance to both city-states and individuals (Cartwright 2013).

[16] Barnard Williams calls such freedom 'metaphysical freedom' (2008: 152).

determined by ultimate fate[17] but still leaving freedom to act within such constraints and thereby be responsible for actions. As will be seen later, these ideas resonate with the compatibilism of John Martin Fischer (2013) and many others in that personal freedom is possible within a world that may be determined.

o Early Greeks: Plato, *Gorgias*

From the epic works of Sophocles and reflections on human freedom within the context of supernatural adventure, human imperfections and tragic heroes, I now turn to Plato's dialogue the *Gorgias*. The introduction to the 2004 Penguin Classics addition of *Gorgias* says 'Plato stands with Socrates and Aristotle as one of the shapers of the whole intellectual tradition of the West' (Plato 2004: 30).[18] Engaging with, what Terence Irwin describes as, 'a puzzling and unsatisfactory dialogue' (1977: 131), Plato's objectives within *Gorgias* will be briefly outlined, including reference to the long, complex and hugely important work within the Platonic corpus, *The Republic*.

The dialogues of Plato discuss basic questions of morality. As Irwin notes at the very beginning of his detailed examination of Plato's moral theory, 'both Socrates and Plato defend some controversial and puzzling answers to these (moral) questions' (1977: 1). For example, claims that the virtues of courage, temperance, piety, justice and wisdom are in the *agent's* self-interest and have intrinsic good distinct from their consequences in action. Socrates and Plato believed such claims were entailed within contemporary conventional beliefs of their fellow citizens and accessible through dialogue and cross-examination. As moral philosophers, Plato and Socrates look critically at moral beliefs and recognise that vital to their project is clear and agreed understanding of fundamental human virtues such as courage and temperance. It is clearly necessary for Socrates and Plato to interrogate the meaning of terms, such as just and unjust, to allow meaningful discussion of whether something *is* just or unjust, or how justice is best accomplished.

[17] There is a distinct and major difference between Fatalism and Determinism. Fatalism claims the inevitability of major events that happen in life such as choice of marriage partner, the place one choses to live and so on, typically expressed by 'it's meant to be'. Whereas, for Determinism it is some form of necessary causality bounded by natural laws that drive and determine outcomes.

[18] Citation of Kindle books will include guide location references where real page numbers are not available.

Further, the nature of virtue is important because whether the concept entails *necessity* to act virtuously will directly affect how freedom within the dialogue is to be understood. Plato's position on virtue is generally more challenging to understand than it may first appear, as Plato's position changes, or better expressed, develops, during later dialogues. A vital point, which cannot be discussed within this Thesis, but must be mentioned is the vital role of Socrates, individual rationality and the notion of dialogue as the pathway that leads to knowledge concerning ethical questions. The importance of this development within moral philosophy cannot be overstated.[19]

Virtue can be expressed as an excellence in performing a particular function. For example, dogs, musical instruments and athletes have or lack virtue depending on whether they succeed or fail to meet what is expected within the role of being a dog, a musical instrument or an athlete. (Following Republic 353b2 - d1 and 335b6 - c2). For Man[20] the situation is clearly different. While expectations about what we should do and how we should behave as human beings are clearly more complex, more important, as human beings we must *decide* what we expect of ourselves and others. The choice is difficult, not least because we do not have a clear, obvious or given function. Is there a common expectation shared by everyone? Socrates claims that a final good, such as happiness, flourishing or living well[21] is the expectation of everyone and authentically conducting life in a way that achieves the final good is to be a virtuous man. Perhaps the most fundamental question is *how* should we live to achieve our expectation of happiness, of flourishing, and so be *wholly* virtuous? The answer is to develop a *virtuous character* by being authentically prudent, just, courageous and adopting temperance.

It is not possible to examine or describe in detail these issues within this Thesis. However, I will briefly describe part of the *Gorgias* dialogue where two alternative and opposed ways of life and the form of freedom they entail are examined. The first,

[19] An excellent description of this time of transition is presented in the television documentary *Socrates Genius of the Ancient World* by historian Bettany Hughes, first broadcast on Wednesday 12th August 2015 BBC Four.

[20] Unless obviously otherwise, gender specifics such as 'Man' and 'Men' are used here and throughout to describe inclusively all human beings.

[21] The term 'happiness' is a common translation of the Greek 'eudaimonia' however, 'human flourishing' has been proposed as a more accurate translation.

articulated most clearly by the character Callicles in the third dialogue, considers the maximisation of pleasure through exercise of power to be the obvious, good and correct basis for conducting our lives. The rival position, advocated by Socrates, is a life concerned with and directed by moral absolutes such as justice and goodness guided by *reason*. In terms of human freedom, Callicles believes true freedom is expressed by unrestricted power to experience *whatever* pleasure is desired; maximum pleasure and desire satisfaction constitutes the good life. Socrates defines and advocates freedom not in terms of unconstrained choice but a concept of freedom embracing 'reasoned choice in line with virtue' (McLoughlin 2012: v).[22]

What is the nature of Socrates' freedom, if action is to be 'in line' with virtue? There seems to be a sense of necessity within such an idea. Chris Emlyn-Jones' introduction to *Gorgias* describes Socrates position clearly, where knowledge of the good necessitates good action. For Socrates

> a necessary precondition of doing right or good is to *know* what actually is right and good, i.e. its nature; moreover, the precondition is not only necessary *but sufficient*: for Plato's Socrates, *to know* what is right in any given situation *is necessarily to do it;* and once you really know what is right and good, you cannot want to do wrong. (added emphsis Plato 2004: 274)

There is, as would be expected, supporting text within the dialogue for Emlyn-Jones' comment concerning knowing what is right and not wanting to do wrong. For example, within the third dialogue Socrates says, '[…] no one does wrong willingly, and that all wrongdoing is involuntary' (Callicles 509). Not wanting to do wrong, even if correct, clearly does not entail wanting to do right or making right actions necessary. Further examination of the text shows various claims and assumptions made by Socrates that develop through dialogue towards a position on these issues.[23] The important conclusions are (i) For Socrates, the craft of justice or any virtue is different from all other crafts in the sense that it is impossible after acquiring the craft of virtue to cease

[22] See also later references to Susan Wolf, *Freedom within Reason* (1993).

[23] For a systematic and comprehensive study of freedom within Plato's dialogues see Siobhán McLoughlin's PhD Thesis, *The Freedom of the Good: A Study of Plato's Ethical Conception of Freedom* (2012), available open access <http://digitalrepository.unm.edu/phil_etds/15>.

being virtuous *and* impossible to cease acting virtuously. (ii) A rational man, in the fullest sense of someone who shows reasoned self-control and is well ordered, wanting the good, *necessarily* embraces the reasoned and rational choice in accordance with virtue and justice. From the perspective of freedom, such a man does what he *wants* to do; there is no experience or feeling of necessity or compulsion but a real and genuine sense of freedom to act in harmony with a *balanced and virtuous nature*.
Dilman captures the point:

> It is in such self-mastery that the person will have achieved autonomy: what he does will be what he wants to do, not what he is forced to do, and he will be wholly behind it. This is Plato's description of how, inevitably living in a world of natural necessity, a man can be nevertheless free. (1999: 29)

There is much that could be questioned, discussed and analysed that unfortunately falls outside the scope of this Thesis. As previously cited, Terence Irwin's *Plato's Moral Theory* (1977) presents a very detailed and systematic analysis of the early and middle dialogues. An important point to note, is that at the heart of the above are the problematic claims that virtue can be an object of knowledge *and* whoever has knowledge of virtue is virtuous *and* necessarily acts virtuously.[24]

Within the context of Plato's Gorgias, a conception of freedom is described where action is necessitated in an important way, (a moral way), yet it is still legitimate to claim such action is free, being a deliberate choice and wholly consistent with what an autonomous agent wants. Rosalind Hursthouse, quoted by Nafsika Athanassoulis in *Virtue Ethics* from the Internet Encyclopedia of Philosophy, summarises much of what has been said so far and makes clear the important point that acting virtuously is *of itself* rewarding and not something desirable simply in terms of good consequences:

> […] virtue is not exercised in opposition to self-interest, but rather is the quintessential component of human flourishing. The good life for humans is the life of virtue and therefore it is in our interest to be virtuous. It is not just that the virtues lead to the good life (e.g. if you are good, you will be rewarded), but

[24] See also Jaakko Hintikka's *Knowledge and the Known*, for relevant discussion of virtue, skill, belief and knowledge (Hintikka 1991: 28).

> rather a virtuous life is the good life because the exercise of our rational capacities and virtue is its own reward. (Athanassoulis 2017)

- Theological related issues within the works of St. Augustine (354-430 CE) and St. Aquinas (1225-1274)

Many see freedom in the context of the Christian faith as particularly problematic, generating over the centuries a vast body of theological and philosophical reflection. The basic problem has been expressed in many ways; One of the clearest and precise is from Linda Zagzebski's article *Foreknowledge and Free Will*:

> For any future act you will perform, if some being infallibly believed in the past that the act would occur, there is nothing you can do now about the fact that he believed what he believed since nobody has any control over past events; nor can you make him mistaken in his belief, given that he is infallible. Therefore, there is nothing you can do now about the fact that he believed in a way that cannot be mistaken that you would do what you will do. But if so, you cannot do otherwise than what he believed you would do. And if you cannot do otherwise, you will not perform the act freely. (2017)

It is a problem for those who believe it is necessary to accept claims that God infallibly knows the entire future, and, that human beings act freely. To deny either of these claims would be a major difficulty for most believers; the notion of foreknowledge is embedded within Christian theology yet without freedom to act otherwise, responsibility, blame and punishment appear unjustified and so contrary to the notion of a supremely just God. There are many responses to this problem; one of the most important is St. Augustine's *De Libero Arbitrio* (On Free Will) and *The City of God*, particularly Book Five. St. Augustine is described as 'a towering figure of medieval philosophy whose authority and thought came to exert a pervasive and enduring influence well into the modern period' (Mendelson 1997). St. Augustine produced a huge body of work during his long life, much of which survives; over one hundred titles, many of which are themselves voluminous and composed over a long time. In addition, there are over two hundred letters and nearly four hundred sermons. St. Augustine's position evolved as Christian scripture became increasingly influential in his life. In terms of the free will and Divine foreknowledge problem, St. Augustine's objective was to reconcile the absolute necessity of God's knowledge, (i.e., if God genuinely knows that x is going to happen, it is impossible for x not to take place, see *De Libero Arbitrio* III.4 and *De Civitate Dei* V.9), with the claim that there can be no moral responsibility unless it is within the agent's power to choose to do other than what in fact was chosen, (see *De Libero Arbitrio* III.3). At the heart of the problem is the notion of necessity that underlies the Greek conception of knowledge; simply expressed, something that is known cannot be false, it is necessarily true,[25] in this case God's knowledge of the future. Seeking compatibility between God's infallible knowledge of both the past and future, and human free will, St. Augustine reflects on the nature of the human will. Based on *City of God*, Book Five, Chapter Ten, Linda Zagzebski expresses St. Augustine's position:

> His argument seems to involve an account of what it is to will. To will just is to act voluntarily and this in turn means to have it within our power. He thinks this is so because our wills can be contrasted with things that are obviously not in our power such as growing old. So, for Augustine any act of will is a voluntary act and a free act. But this does not preclude the necessity claimed by the determinist […]. (Zagzebski 1985: 280)

In addition to Chapter Ten, Whether Our Wills are Ruled by Necessity, other chapters from Book Five reflect the deep struggle St. Augustine experienced with this problem and provide further insight into St. Augustine's position: Chapter Eight; […] the Connection of Causes that Depend on the Will of God and Chapter Nine; Concerning the Foreknowledge of God and the Freewill of Man […].

Discussing fate, Chapter Eight begins with the interesting and relevant statement that essentially it is acceptable to describe fate as '[…] the whole train and connection of causes which makes everything become what it does become', such connection of causes being attributed to 'the will and power of God most high, who is most rightly and most truly believed to know all things before they come to pass and leave nothing unordained' (1871: 209). This definition of fate contrasts with future outcomes being decided by time of birth and the position of the stars, something that St. Augustine would clearly reject. The remaining text of Chapter Eight presents more questions than answers. After saying

[25] My description of St. Augustine's objective generally follows Michael Mendelson, *The Stanford Encyclopedia of Philosophy* (1997).

that God leaves nothing 'unordained' and 'from whom are all powers' there follows within the same sentence the statement 'although the wills of all are *not* from Him'. St. Augustine continues, '[…] it is *chiefly* the will of God most high, whose power extends itself irresistibly through all things […]' (added emphasis). The Chapter concludes with lines from Seneca and Homer that speak of fate leading the willing but dragging the unwilling, suggesting it is ultimately impossible to resist fate, however unwilling. This appears counter to the earlier statement that the wills of all are *not* from Him suggesting it could be possible to avoid the hand of fate. Chapters Nine and Ten seek to clarify some of this confusing and apparently inconsistent set of statements and claims, defending the compatibility of divine foreknowledge and freedom. As will be seen, the claims and arguments are not straightforward. For example, Chapter Nine may be summarised, informally, as follows:

From Chapter Nine:

- God has infallible foreknowledge, *and* we have free will. St. Augustine tells us not to be fearful about this matter if it turns out that '[…] we do not do by will that which we do by will because He, whose knowledge is infallible, foreknew that we would do it' (1871: 211).
- 'But it does not follow that, though there is for God a certain order of all causes, there must therefore be nothing depending on the free exercise of our own wills, for *our wills themselves are included in that order of causes* which is certain to God, and is embraced by His foreknowledge, for human wills are also causes of human actions; and He who foreknew all the causes of things would certainly among those causes not have been ignorant of our wills' (added emphasis 1871: 213).

This quotation appears to capture St. Augustine's compatibilist[26] position, that human free will exists *and* is in some sense bounded or included within divine foreknowledge. The meaning of 'bounded or included within' needs clarification if compatibility between the apparently opposing claims of foreknowledge and freedom is to be defended convincingly.

[26] See Chapter 2 for description of compatibilist and other positions on free will and determinism.

What can be concluded from this challenging text? For St. Augustine, human freedom exists within an all-encompassing divine foreknowledge. God is responsible for creating and sustaining the whole fabric of human existence including our freedom to act; a legitimate freedom that brings responsibility while sustained by and transparent to the divine.[27]

For the will to be undetermined and free, and actions to be foreknown and necessary remains problematic. Subsequently, philosophers and theologians have developed strategies that try to reconcile these disparate requirements, for example, resetting what is understood to be the nature of God's knowledge by reducing what is known[28] or arguing that God does not exist in time in any sense that is conceivable by human beings, hence our human notion of foreknowledge has no actual meaning for the Divine. On this view, because God comprehends everything that has happened and will happen simultaneously it is suggested that God cannot be said to know things in advance.[29] It is not surprising there are numerous problems and issues concerning the idea of God's knowledge given the problematic nature of each individual concept.[30] That said, it is very clear that free will and considerations of blame and responsibility *are vital* within traditional Christian belief as the implications for believers extend beyond this short life to all eternity.

To conclude this brief reflection on human freedom within a theological context the ideas of St. Aquinas (1225-1274) will be described. The text to be considered is the first part of *The Summa Theologica*, Question 83; Of Free Will (consisting of 4 points of enquiry) (1947) written by St. Thomas Aquinas, one of the Catholic Church's greatest theologians and philosophers, between 1265 and 1274 C.E. It is the first point or article of enquiry that is most relevant; Whether man has Free Will? Five objections to Man

[27] St. Augustine's *De Libero Arbitrio* (On Free Will) is not considered here. It is unfortunately impossible to respond appropriately to such a work within the confines of an overview. See for example Russell Danesh Hemati's Thesis *St. Augustine's Solution to the Problem of Theological Fatalism* (2010) for detailed consideration of Human freedom within St. Augustine.

[28] See Richard Swinburne *The Coherence of Theism* (1993).

[29] See Eleonore Stump and Norman Kretzmann *Eternity* (1981).

[30] See Edward Wierenga *Omniscience* (2011).

possessing free will are noted and supported by Biblical quotations. St. Aquinas makes a general response to these objections with a Biblical quotation that supports his position; 'God made man from the beginning and left him in the hand of his own counsel' *Ecclesiasticus* 15:14 (King James Version). St. Aquinas then makes a more substantial general reply pointing out that without free will, reward and punishment would be in vain and continues by describing how some things act without judgment, such as a stone moving downwards, some act from judgment but not free judgment, such as animals, and lastly, how man acts from judgment because by his apprehensive power he judges that something should be avoided or sought. This judgment, in the case of some particular act, is not from instinct but from some act of comparison utilising reason, therefore man acts from free judgment *and* retains the power of being inclined to various things. Particular operations are contingent, therefore in such matters the judgment of reason may follow opposite courses and is not determinate to one. And forasmuch as man is rational is it necessary that man have a free will (my summary of St. Aquinas 1947: Question 83).

Having made this general response, the five objections are addressed individually. These make important claims that draw discussion back to the determining role of God, not framed in terms of foreknowledge directly but in terms of causation. For example, St. Aquinas says in reply to objection two, Man's free will is not sufficient *on its own* but must be moved and helped by God. This is clarified in reply to the third objection, where free will although the cause of its own movement is not necessarily the first cause. God is the first cause, who moves causes both natural and voluntary. By moving natural causes He does not prevent their acts from being natural, so by moving voluntary causes He does not deprive their actions of being voluntary and subject to responsibility[31] (my summary from 1947: Question 83). How can moving voluntary causes and not depriving their actions of being voluntary be understood? Understanding is possible for St. Aquinas by considering God's relation to time, mentioned previously. It is beyond the

[31] St. Aquinas is not suggesting joint responsibility for such actions, as made clear in the following; 'It is also apparent that the same effect is not attributed to a natural cause and to divine power in such a way that it is partly done by God, and partly by the natural agent; rather it is wholly done by both, according to a different way, just as the same effect is wholly attributed to the instrument and also wholly to the principal agent' (*Summa Contra Gentiles* III.70.8 translated by Vernon J. Bourke).

scope of this section to provide a comparative analysis of God's causation and Man's causation, but in essence it is claimed that the nature of God with respect to time is such that human causation can be truly free within an all embracing, all knowing God. St. Aquinas describes God's knowledge and temporal status at length in *Summa Contra Gentiles* I.64-71, (particularly Chapter 66.7 and 67). The essential claim is that God's knowledge relates equally to every moment of time that from our perspective is expressed in terms of past, present and future. Therefore, the term foreknowledge is at best misleading as it assigns a position in time to God's knowledge of some contingent event that will occur in our future. Perhaps it is better to say that God simply knows rather than foreknows. From God's perspective, certain and unchangeable knowledge of future (for us) events is possible even when in themselves they are contingent and liable to change.[32]

Is this a convincing argument for compatibility of foreknowledge and human freedom? It is the case with all the perspectives described within this Chapter that much more could and perhaps should have been said. Here, the situation is particularly acute given the scale of works by the hugely prolific, influential and historically prominent figures St. Augustine and St. Aquinas. Reassessing the relationship between time and the divine such that God has total knowledge of (for us) the past, present and future in one timeless gaze appears a positive and plausible suggestion in terms of trying to resolve the dichotomy of foreknowledge and freedom. That said, the challenge of foreknowledge to our sense of freedom is still hugely demanding to resolve satisfactorily, with obvious implications for our understanding of responsibility, praise and blame. At the centre of this issue is one of the most difficult and intractable concepts in philosophy, the nature of God.[33] Ascribing to a divine being, existing in a profoundly different temporal mode, concepts such as knowledge and causation, each a difficult and contentious concept, is obviously extremely challenging and problematic. Such is the ontological difference between man and God that *ultimately*, resolving questions concerning our notion of foreknowledge is surely impossible?

[32] There is a sense of incongruity when describing something 'from God's perspective'.

[33] For example, expression of this idea is made by the character Demea from David Hume's *Dialogues Concerning Natural Religion* 1779; God exists but has a nature beyond our ability to understand.

- Free Will within Descartes (1596-1650)

John Cottingham describes René Descartes, the 'Father of modern philosophy', as '[…] occupy(ing) a pivotal role in the transition from the widely accepted scholastic view of science to its complete rejection, and the emergence of what we think of as modern scientific methodology' (1992: 12). (The context of this quotation may be found in Brien Brockbank's Paper *Descartes and Scholasticism: An Analysis* (2019)). To begin consideration of Descartes' complex and changing position on human freedom I will refer to two quotations from perhaps his most well-known work, published in 1641 in Latin, *Meditations on First Philosophy*. The quotations are from *Meditation IV, Of the True and the False,* a meditation whose main aim is to answer the question, if God is perfectly good and the source of everything, then how is it possible that error and falsehood exist within creation? The answer is essentially that mistakes in judgment are *our* fault because we do not restrict our will to clear and distinct ideas. Attention will focus on freedom within this meditation, but first it is necessary to engage with one of Descartes' most fundamental objectives, to set up a solid metaphysical basis for knowledge whereby reason, when functioning correctly provides truths, guaranteed by a non-deceiving God. 'Clear and distinct ideas - the tools of reason - may confidently be used: God guarantees their reliability' (Berman 2004: 2):

> God exists, and that my existence depends entirely on Him in every moment of my life — I do not think that the human mind is capable of knowing anything with more evidence and certitude. And it seems to me that I now have before me a road which will lead us from the contemplation of the true God (in whom all the treasures of science and wisdom are contained) to the knowledge of the other objects of the universe. For, first of all, I recognize it to be impossible that He should ever deceive me; for in all fraud and deception some imperfection is to be found, and although it may appear that the power of deception is a mark of subtlety or power, yet the desire to deceive without doubt testifies to malice or feebleness, and accordingly cannot be found in God. (Descartes 1911: 19)

The second quotation considers the faculty of the will:

> For although the power of will is incomparably greater in God than in me […] it nevertheless does not seem to me greater if I consider it formally and precisely in itself: for the faculty of will consists alone in our having the power of choosing

> to do a thing or choosing not to do it (that is, to affirm or deny, to pursue or to shun it), or rather it consists alone in the fact that in order to affirm or deny, pursue or shun those things placed before us by the understanding, we act so that we are unconscious that any outside force constrains us in doing so. For in order that I should be free it is not necessary that I should be indifferent as to the choice of one or the other of two contraries; but contrariwise the more I lean to the one — whether I recognize clearly that the reasons of the good and true are to be found in it, or whether God so disposes my inward thought — the more freely do I choose and embrace it. And undoubtedly both divine grace and natural knowledge, far from diminishing my liberty, rather increase it and strengthen it. Hence this indifference which I feel, when I am not swayed to one side rather than to the other by lack of reason, is the lowest grade of liberty, and rather evinces a lack or negation in knowledge than a perfection of will: for if I always recognized clearly what was true and good, I should never have trouble in deliberating as to what judgment or choice I should make, and then I should be entirely free without ever being indifferent. (1911: 21)

A long quotation, but one that resists summary, particularly points concerning indifference and freedom even when God disposes inward thought towards a particular action. There is much that may be, and has been, said about this text; the key points are as follows. God can produce within man inclinations or a disposition that gives actions at least some degree of predictability, ('impossible that He should ever deceive me') but are not absolutely determined. Clear and distinct ideas enable freedom of spontaneity; in the limit, this is perfect freedom and a determination towards the good and true. Whether this is absolute determination is a controversial point. Confused or obscure ideas bring about freedom of indifference, an imperfect freedom, ('the lowest grade of liberty'), that is not determined.[34] There is a problem concerning Descartes' claim that Man and God's wills are essentially the same except the latter being incomparably greater. Part of God's perfection is being completely undetermined, having profound indifference in terms of action. For Man, the situation is just the opposite, when God so disposes our inward thoughts to the good the more freely they are chosen. Tad Schmaltz expresses the point, 'In contrast to the case of God, then, the perfection of our freedom

[34] There is much debate concerning freedom of indifference and freedom of spontaneity; it is beyond my present scope to explore these issues but detailed analysis may be found within Tad Schmaltz's fascinating book, *Descartes on Causation* (2008).

consists in the determination to the true (and the good) rather than in an indifference to action' (Schmaltz 2008: 196).

Further insight into Descartes position on human freedom is found within the huge exposition of his system completed in Latin in 1644, the *Principia Philosophiae,* (or *Principles of Philosophy*). From Part One, the following articles are relevant:

> Article 37. The highest perfection of man is that he acts freely or voluntarily, and that is what makes him deserve praise or blame.
>
> […] When we embrace something true, that is much more to our credit if we do it voluntarily than it would be if we couldn't help embracing it.
>
> Article 39. It is self-evident that there is free will.
> There's freedom in our will, and we often have the power to give or withhold our assent at will — that is so obvious that it must be regarded as one of the first and most common notions that are innate in us. […].
>
> Article 40. It is also certain that everything was preordained by God.
>
> Article 41. How to reconcile the freedom of our will with divine preordination.
> But we will get out of these difficulties, (trying to reconcile divine preordination with the freedom of our will), if we bear in mind that our mind is finite, and that God has infinite power by which he not only knew from eternity everything that was or could be going to happen, but also willed it and preordained it. We can know enough about this power to perceive vividly and clearly that God has it; but we cannot get our minds around it well enough to see how it leaves men's free actions undetermined. As for our own liberty — our ability at a given moment to go this way or that — we are so intimately aware of this aspect of our nature that we see it as clearly and comprehend it as fully as we do anything. When something is as intimately and securely grasped as that, it would be ridiculous to doubt it just because we don't grasp something else — namely its relation to God's powers of knowledge — that we know must by its very nature be beyond our comprehension. (Descartes 2012: 9-10)

Because of the quantity of text only part of Articles 37 and 39 are presented. Article 41 has been given in full, as it is particularly relevant. In one sense the above is quite clear. The claims are robustly made but Article 39 and 40 lead to a familiar difficulty, (trying to reconcile divine preordination with the freedom of our will), that Article 41 is intended to resolve. It is worth noting that unlike *Meditation IV* there are no claims that determination, even if understood loosely as an inclination, enables a higher form of freedom. Article 37 seems to claim that real freedom precludes determination of any sort, giving actions of automata as an example of determined behaviour that clearly cannot be subject to praise or blame. Returning to Article 41, it is suggested by Brian Collins in his Paper *Adding Substance to the Debate: Descartes on Freedom of the Will* (2013) that Descartes' position is illuminated by considering an analogy between divine and created substance and divine will and human will:

> God is a substance in the primary sense because God is completely independent (i.e., not dependent on anything). Humans are substances in a secondary sense because we are independent of all substances other than God. God is free in the primary sense because God is completely undetermined. Humans are free in a secondary sense because our will is only internally determined by our recognition of truth and goodness (which is created and dependent on God). (2013: 232)

Collins continues with clarity, noting that 'for human freedom, the degree of freedom increases as the will acts with greater ease (i.e., more spontaneously/voluntarily). When the will acts with more spontaneity it is following a more clearly perceived good' (2013: 233). Is Collins' suggested comparison, (substance in the primary or secondary sense), helpful and does Descartes' claim that our liberty is 'intimately and securely grasped … and it would be ridiculous to doubt it just because we don't grasp something else' bring any kind of satisfactory closure, resolution or explanation? I believe that it does not, but there are more reflections by Descartes on these issues to be considered.

In a letter to the Jesuit Denis Mesland of 9th February 1645, well known in the ongoing controversy concerning Descartes' position as Compatibilist or Libertarian,[35] Descartes again discusses the nature of the will. Descartes describes a freedom of indifference where the content of what is chosen is not important. In other words, it is possible to choose arbitrarily among possibilities in quite an indifferent way. While there seems no deep inconsistency with the *Meditations* where Descartes writes, 'For it is always open to us to hold back from pursuing a clearly known good, or from admitting a clearly perceived truth, provided that we consider it a good thing to demonstrate the freedom

[35] Responses to the free will problem in terms of Libertarianism, Compatibilism and Incompatibilism will be described at some length in Chapter 2 of this Thesis.

of our will by so doing' (Lennon 2013: 225)[36] it is difficult to imagine circumstances that would lead an agent to consider it a good thing to make such a demonstration. In addition to the mysterious demonstration of freedom, it is difficult to imagine why one would want to hold back from experiencing higher freedom by pursuing a guaranteed and clear good, having its origin in God. Detailed analysis of this letter is available[37] and there are further Works within Descartes' body of writing that continue to refine and develop his position on the freedom of the will and the closely related issues of God's knowledge and causation. *The Passions of the Soul* (1650) and correspondence with Elisabeth of the Palatinate (1618 – 1680), (also known as Elisabeth of Bohemia, Princess Elisabeth of the Palatinate or Princess-Abbess of Herford Abbey), are further examples of substantial insights into Descartes' thought that cannot be considered in this brief outline. For Descartes, free will is self-evident and in an imperfect sense may be likened controversially with the freedom of God. Human freedom exists in its highest form when choosing the good in harmony with God's desire and influence. Such influence could, in its strongest interpretation, be described as determination, thereby bringing together the apparently disparate requirements of human freedom within a necessarily determined world. It is interesting to note the concept of freedom just described, where action is necessitated and yet also free by alignment of the will with the divine or the good, although separated by time, culture and purpose, has appeared previously in various forms from the beginning of this brief survey.

- Free Will within Hume (1711-1776)

David Hume is 'generally regarded as one of the most important philosophers to write in English' (Morris and Brown 2014), and 'widely recognized as providing the most influential statement of the "compatibilist" position in the free will debate' (Russell

[36] This quotation is taken from the full text of Descartes' letter to Mesland reprinted in Thomas M. Lennon's Paper, *Descartes's Supposed Libertarianism* (2013). The Paper discusses in detail the controversy concerning Descartes' Libertarian or Compatibilist position.

[37] Detailed analysis of the entire letter can be found in C.P.Ragland *Descartes on Degrees of Freedom: A Close Look at a Key Text* (2013).

2014).[38] (A position claiming freedom and moral responsibility can be reconciled with the truth of (causal) determinism). Hume's work on free will is contained within *Of Liberty and Necessity* published in *A Treatise of Human Nature* 1738-40 (Book 2, Part 3, Sections 1 and 2) and later, in slightly amended form in *Enquiry Concerning Human Understanding* (Section 8, Parts 1 and 2, *Liberty and Necessity*). To provide an outline of Hume's ideas, I will look at *Enquiry Concerning Human Understanding* Section 8, and for better understanding also briefly mention some of Hume's ideas in the preceding Section 7, *The Idea of Necessary Connection.*

Hume believed lack of progress in resolving metaphysical issues was due to ambiguity of key terms and obscure ideas. Separate groups could not agree because although using the same terms there were different assumptions about the meaning of those terms. Hume believed the debate about free will was protracted because of this problem and resolved to clarify the terms used and effectively end the debate. Hume believed that clarifying key terms such as necessity and liberty would allow everyone to agree that human beings are free, *and* actions follow necessarily, and by so agreeing end the long running debate. First, from Section 7, Hume addresses causation:

> We are never able to discover any power or necessary connection, any quality that ties the effect to the cause and makes it an infallible consequence of it. All we find is that the one event does in fact follow the other. (2004: 31)

In other words, it is claimed the idea of necessity and causation arises entirely from the uniformity we see in the operations of nature where similar items are constantly conjoined. The mind is determined by custom to infer the one from the appearance of the other. The necessity that we ascribe to matter is based on the constant conjunction of similar objects and the consequent inference from one to the other. Apart from these we have no notion of necessity or connection. This idea is extended from the material world to the world of human action. To give up our clear sense of personal freedom seems wrong, but to be consistent these ideas must surely be extended to human action:

[38] See also 'Hume's Lengthy Digression': Free Will in the Treatise, *Hume's Treatise: A Critical Guide* (Russell 2015b).

> The same (one event following another) holds for the influence of mind on body: the mind wills, and then the body moves, and we observe both events; but we don't observe — and can't even conceive — the tie that binds the volition to the motion. (2004: 36)

Hume argues that our life experience and history generally support this claim, that to study history is to discover universal principles of human nature. The essential claim is that the mysterious process of constant conjunction legitimately describes material processes in nature *and* the relationship between motives and voluntary human actions. This seems counter to experience; it feels from the inside that human actions are *not* universally predictable in the sense that other events in nature are predictable and human beings would be diminished by such predictability. Hume partly accepts this, suggesting differences in character are responsible at a fine level of detail for different outcomes under the same initial conditions. Further, Hume suggests (from Section 8, Part One) because '[…] a human body is a mighty complicated machine and many secret powers lurk in it that we have no hope of understanding' (2004: 43), some causal factors are effectively hidden, making it only appear as if there are variations in the effect of apparently identical causes. There seems to be a problem, with endless variation in human behaviour driving ever more complex and possibly hidden preconditions such that the claim of constant conjunction seems to be getting lost in ever increasing detail. With material objects this problem does not seem to exist, with relationships of constant conjunction much more apparent. Familiarity and dependability of constant conjunction within the material world, such as released objects falling, does not diminish the mysterious nature of the conjunction of these events. Hume describes constant conjunction in the material world and in the context of human action as 'entirely incomprehensible' and something 'we don't observe and can't even conceive' (2004: 36). However, although mysterious, for Hume there is nothing in addition within the causal process for material or human related causation.

How is human freedom understood within Hume's universal conjunction model of cause and effect? Liberty means a power of acting or not acting according to the determinations of the will. 'This hypothetical liberty - hypothetical because it concerns what we may do if we so choose - is universally agreed to belong to everyone who isn't a prisoner and in chains' (Hume 2004: 48). It is claimed that for morality to have foundation the agent's character must have the power to produce sentiments, (feelings and opinions), which have a constant effect in producing actions. There is constancy and predictability in the way character drives sentiments and actions. In Hume's words, 'It seems almost impossible, therefore, to engage either in learning or in action of any kind without acknowledging the doctrine of necessity, and this inference from motives to voluntary actions, from characters to conduct' (2004: 45). Such a doctrine of necessity and liberty is essential to Hume's notion of morality.

For Hume, necessity, in the sense of cause-and-effect relationships that are observed as objects constantly conjoined, is true within the material world and within human mental and physical activity. Liberty to act is possible, whereby morally relevant actions originate from within the agent, flowing *from* the nature of the agent's character. (It will be seen later, this is a problematic notion due to, for example, the role of luck in the formation *of* an agent's character). Thus, existence of morality together with associated praise or blame is claimed to be possible. For Hume, both necessity and liberty exist and are compatible, being the basis of our experience of morality.[39] Further reflection quickly shows issues for consideration. As mentioned, actions originate from the agent; I do what I want to do, driven by character, plans, objectives and so on. If action is *determined* by character, there seems to be only one possibility; at the moment of choosing there is one choice based on what it is that I want. It seems misleading perhaps to describe such a choice as free. William James (1842-1910) discussed such difficulties during a lecture[40] to Harvard Divinity School students, calling compatibilism a 'quagmire of evasion'. These issues will continue to be explored in the next section while describing the ideas of Kant on free will.

[39] With reference to previous discussions of Divine foreknowledge, it is interesting to note that David Hume claims 'It has so far been found to be beyond the powers of philosophy to reconcile the indifference and contingency of human actions (so that men could have acted differently from how they did act) with God's foreknowledge of them, or to defend God's absolute decrees and yet clear him of the accusation that he is the author of sin' (2004: 52).

[40] *The Dilemma of Determinism* was delivered as an address to Harvard Divinity School students in Divinity Hall on March 13th, 1884 at 7:30 pm and published in the Unitarian Review September 1884.

- Free Will within Kant (1724-1804)

Kant's importance in the history of Philosophy is immense. Bryan Magee describes Kant as '… the man most widely regarded as the greatest philosopher since the ancient Greeks …' (1987: 170). In conversation with Magee, Sir Geoffrey Warnock suggests the origin of the problems Kant endeavoured to solve was 'the apparent conflict between the findings of the physical sciences in his day and our fundamental ethical and religious convictions' (Magee 1987: 171). Specifically, conflict between the ascendancy of determinism within the physical sciences and necessary belief in alternative possibilities within the realm of human behaviour that allows responsibility to be assigned to our actions.[41] Magee and Warnock agree that as a builder of systems, it is difficult to explain or describe one aspect of Kant's thought in isolation; 'an immense range of views fit together in a systematic and comprehensive way' (1987: 171). However, the aim here is to provide a brief outline of Kant and free will, and I will endeavour to begin with minimum preliminaries. I will refer to *The Critique of Practical Reason* (1889), translated by Thomas Kingsmill Abbott: (i) Section I: Fundamental Principles of the Metaphysic of Morals, Third Section, and, (ii) Section II: Critical Examination of Practical Reason, First Part, Book 1, Chapter 3. On Kant's view, moral praise, blame, the very legitimacy of moral appraisal, presupposes that an agent is able to do otherwise (1889: 190). Kant makes a clear and vital distinction between the world as appearance, (the world of the sciences, where physical determinism rules), something that is the object of our experience, and the world of things in themselves, (a world that includes, for example, free will, right and wrong and the soul) (1889: 70). As Warnock notes, to position right and wrong outside of the world of experience raises the obvious and unresolved problem of how the will and moral thought can actually 'make any difference' or connect with the world of experience. (Magee 1987: 183). Kant's essential method is to place the causality of science, of matter in motion, in the world of appearances that exists within the context of time and space, and place freedom within the world of things in themselves, outside of time, so immune from any notion of causality and determinism.

[41] In a typically clear and incisive manner, Bryan Magee describes the (free will) problem; 'So the problem is: how, in a universe in which the motions of all matter are governed by scientific laws, can any of the motions of those material objects which are human bodies be governed by free will?' (1987: 172).

Section I: Fundamental Principles of the Metaphysic of Morals, Third Section, begins by describing the will as a kind of causality that is possessed by rational living beings[42] and freedom as a property of this causality, independent of determination by 'foreign causes' (1889: 65). By contrast, physical necessity is the property of causality relating to irrational beings and *is* determined by 'foreign causes'. Defining freedom in terms of what it is not, (not dependent on foreign causes), is considered unfruitful in terms of revealing its essence. Freedom is not lawless, except in terms of physical laws, because any form of lawlessness will make free will 'an absurdity'. It is claimed that the will is autonomous, that is, it is a law-to-itself. In the text that follows, Kant seeks to express and develop the relationship between freedom, autonomy, morality, law, rationality and the categorical imperative, ('the idea of the will of every rational being as one that legislates universal law'[43] (R. Johnson and Cureton 2016)). This proves to be ultimately problematic, with Kant commenting that further work is required. The difficulty that Kant recognises is 'it must be freely admitted that there is a sort of circle here from which it seems impossible to escape' (1889: 69). However, escape is thought to be possible as 'one resource remains'.

The 'remaining resource' is the possibility of seeing ourselves from two different points of view. First, as a subject of experience mediated by the senses, (passive and originating from without), and subject to laws of nature. Second, 'as belonging to the intelligible world, *under laws which being independent of nature* have their foundation not in experience but in reason alone', (added emphasis 1889: 72). In other words, to be independent from the determining causes of the sensible world *is* freedom. Kant believes the circle has been broken; 'Now the suspicion is removed which we raised above, that there was a latent circle involved in our reasoning' (1889: 72). Kant expresses the whole point as follows, a long quotation but one that captures in one elegant statement the essence of Kant's position:

[42] The translator, (Thomas Kimsmill Abbott), uses the term 'living being' rather than specifically human being. Presumably, in Kant's view, it is possession of particular attributes such as autonomy that is decisive. See also Kant: *Anthropology from a Pragmatic Point of View* published in 1798 for further general discussion.

[43] There are several formulations of the Categorical Imperative. This formulation is usually described as the Autonomy Formulation (R. Johnson and Cureton 2016).

> […] hence, he (Man) has two points of view from which he can regard him-self, and recognise laws of the exercise of his faculties, and consequently of all his actions: first, so far as he belongs to the world of sense, he finds himself subject to laws of nature (heteronomy); secondly, as belonging to the intelligible world, under laws which being independent on nature have their foundation not in experience but in reason alone.
>
> As a rational being, and consequently belonging to the intelligible world, man can never conceive the causality of his own will otherwise than on condition of the idea of freedom, for independence on the determining causes of the sensible world, (an independence which Reason must always ascribe to itself), is freedom. Now the idea of freedom is inseparably connected with the conception of autonomy, and this again with the universal principle of morality which is ideally the foundation of all actions of rational beings, just as the law of nature is of all phenomena.
>
> Now the suspicion is removed which we raised above, that there was a latent circle involved in our reasoning from freedom to autonomy, and from this to the moral law […]. (1889: 72)

Kant continues, trying to reconcile these two aspects of man; when man is considered free and when man is considered part of nature. In fact, not just reconciliation is sought but the nature of the *necessity* of such unity. My reading of Kant does not ultimately identify a satisfactory conclusion, in fact Kant ends Section I: Fundamental Principles of the Metaphysic of Morals, Third Section, by admitting the limitations of what has been shown, but not in an apologetic tone, noting that sometimes an end stop is reached at the 'very limits of human reason' (1889: 84).

From Section II: Critical Examination of Practical Reason, First Part, Book 1, Chapter 3, Kant develops the significance of time in understanding freedom and determination:

> The notion of causality as physical necessity, in opposition to the same notion as freedom, concerns only the existence of things so far as it is determinable in time, and, consequently, as phenomena, in opposition to their causality as things in themselves. (1889: 188)

Kant continues to argue that if freedom is to be 'saved' then while 'a thing' subject to causation and determinism exists in time and the world of appearance, the attribute of freedom belongs to 'the being' as a thing-in-itself. Freedom does not exist just because the determining physical cause is *within* the person that acts, this is the freedom of the clock and arguments to the effect that such freedom is sufficient are a 'wretched subterfuge' (1889: 189). Freedom in the full and legitimate sense is the necessary basis of moral law and 'imputation', (to enter in the account[44]), to be responsible in a morally important sense.

Importantly, Kant tries to explain and clarify the notion of responsibility *Critique of Practical Reason* (1889: 191). 'A subject' is conscious of himself as being within the world of appearance *and* as a thing-in-himself, not subject to time conditions, determined only by laws that are self-given using reason and the necessary synthesis of moral law and freedom. In this mode of existence nothing is antecedent in a freedom limiting sense to the 'determination of his will'. *Every* action, *including* what Kant describes as 'the whole series of his existence as a sensible being', is within the consciousness of his 'supersensible existence' and the result of *his causality* as a noumenon, (as a thing-in-himself[45]). Kant continues, saying that an agent could have done otherwise because although actions are determined and necessary when seen looking back into the appearance of the past, there is in a mysterious sense a 'single phenomenon' of character, whereby freely created and ongoing actions *of the agent* originate from the intelligible mode of the agent's existence that includes, it will be recalled, 'the whole series of his existence as a sensible being'. Such self-creation gives the agent meaningful responsibility with attendant praise and blame. This is difficult to grasp; the difficulty is largely caused by the notion of two modes of existence that seem in conflict with respect to freedom and determinism. While the idea of freedom existing within a non-causal world is scarcely graspable as an idea, it is the relationship, the interaction of the deterministic world of appearance and the intelligible world, that remains mysterious and incomprehensible.

[44] Origin: Late Middle English; from Old French imputer, from Latin imputare 'enter in the account', from in- 'in, towards' + putare 'reckon'.

[45] *The Stanford Encyclopedia of Philosophy* succinctly describes the 'noumenal self' as an uncaused cause outside of time, which therefore is not subject to the deterministic laws of nature in accordance with which our understanding constructs experience (Rohlf 2016).

There are many questions that quickly emerge from the above description. For example, in addition to the issue of the interaction of the deterministic world of appearance and the intelligible world, how do multiple transcendentally free agents interact? How are my free actions integrated into the experience that your understanding constructs? Even though there are problems like these, Kant holds that we can make sense of moral appraisal and responsibility only by thinking about human freedom in this way; it is the only way to prevent natural necessity from undermining both (following Johnson and Cureton 2016).

In summary, the above does not scratch the surface of the depth and complexity of Kant's response to free will, or more accurately free will as an element within an interlocking and all-embracing explanatory system. However, some insight into his position, in terms of bringing together key elements of determinism, freedom and responsibility, has been suggested.

- Human freedom within Freud (1856-1939) and Sartre (1905-1980).

That philosophical reflection on free will develops and evolves is plainly seen within the work of Kant and Freud. Freud's thinking on free will is clearly described by Alfred Tauber, *Freud, the Reluctant Philosopher* (2010). Chapter Four, The Paradox of Freedom, begins with a summary of the essential claim:

> Basically, Freud divided the mind between the unconscious grounded in the biological and thus subject to some natural causation, and a rational faculty, which lodges itself in consciousness and exists independent of natural cause. The critical distinction resides in Freud's acceptance, as a psychologist, of a functional mind-body dualism, and in the higher functions of the mind, he places the repository of interpretative reason. This is basically a Kantian construction, where-by reason assumes an independent character that allows for a detached scrutiny of the natural world. Beyond this epistemological partitioning, Freud further followed Kant in assigning the scrutinizing ability of a rational self-consciousness the basis of choice and moral reckoning. Consequently, the epistemology leads to a moral philosophy. Freud's debt to Kant thus centres on the dialectical interplay of each domain with the other, and in the end a 'moral-epistemology' emerges. (2010: 116)

A long quotation, but one that captures succinctly a large measure of Freud's position *and* its link with Kant. There is more to say about Freud and his debt to Kant that will further illuminate Freud as 'the reluctant philosopher'. Essentially, Freud's concept of unconscious mental activity 'appears to us … as an extension of the corrections undertaken by Kant of our views on external perception' (from Freud, *The Unconscious* (1915: 121), quoted by Tauber (2010: 117). As Kant argued that our perceptions are subjectively conditioned, not identical with the object that is perceived, so Freud warns us not to equate perceptions by means of consciousness with the unconscious mental processes that are their objects. Freud likened Kant's noumenal self with the psychoanalytic unconscious, neither being *directly* perceived or knowable.

While these connections between Kant and Freud are interesting, what is the relevance for free will in Freud's adoption of fundamental Kantian ideas upon which 'the entire psychoanalytic edifice' (Tauber 2010: 122) was built? Importantly, while Freud believed that the unconscious was unknowable in a way that mirrored Kant's noumenal self, there is a difference in position regarding 'internal' and 'external' perception:

> Like the physical, the psychical is not necessarily in reality what it appears to us to be. We shall be glad to learn, however, that the correction of internal perception will turn out *not* to offer such great difficulties as the correction of external perception - that *internal objects are less unknowable than the external world*. (added emphasis Freud 1915: 121)

That internal objects are 'less unknowable' is based on the claim that reason has a special standing, having the capacity, through *psychoanalysis*, to release from acting in accordance with desires, (unconscious forces), and act autonomously in accordance with reason and moral duty. Note that moral responsibility is the normal and desirable state of existence. As described earlier, reason is independent of the natural world of appearances and causation, it is free and autonomous, directing human beings within a moral landscape.

Freud's enterprise, from the beginning, was built on two fundamentally opposed metaphysical positions. First, humans are determined, allowing Freud's work to be grounded within science and so provide naturalistic explanations governed by naturalistic causes. Second, that humans are free, exercising free choice, independent of natural causation. For Kant, opposition is resolved 'on the basis of reason's standing and the arch-precept of human freedom leading to moral responsibility: 'The entire Kantian philosophical edifice hangs together on these intimate linkages … and while Freud does not address the issue head-on, he basically accepts this construction *and*, with it, assumptions supportive to his own agenda' (Tauber 2010: 140). Freud does not attempt to resolve the free will – determinism paradox and this is a problem because the collision between the determinism exerted by unconscious psychical forces - constituted by naturalistic causation - opposes the other, free will, which Freud often refers to as an illusion. Thus, the structure of Freudian psychoanalysis sits on an unresolved conflict. However, as Dilman points out, 'if Freud was really a strict determinist a priori, if the divisions he attributed to personality and mind were immutable structures, there would be no question of such psychological liberation' (1999: 174). I believe it can be claimed that while Freud tried for good reasons to ground his work upon scientific methods, he could ultimately be forgiven for not resolving the historically intractable free will – determinism paradox that his methodology draw attention to, (while producing and inspiring significant scholarship concerning psychotherapy, science, philosophy, literature, literary criticism and Feminism). From Freud's point of view his methods *worked*, psychoanalysis as a therapy yielded desired outcomes while resolution of what Tauber describes as 'the stark determinism of his science' remained unresolved, put 'to the side as he steadfastly pursued human freedom and self-fulfilment' (2010: 144). Engaging Kant as the philosophical basis upon which to build a theoretical and therapeutic enterprise is certainly not putting issues of determinism too far 'to the side'.

Turning finally to Jean-Paul Sartre, I will briefly outline Sartre's ideas concerning human freedom. Summarising Sartre's ideas is far from straight forward as his initial concept of human freedom meshes with a difficult and complex ontology that later evolves, or certainly changes, into what is often described as a material conception of freedom. Initially freedom is expressed in terms of the ability of consciousness to transcend the material realm; ' … the permanent possibility … of wrenching itself away from its past so as to be able to consider it (it's past) in the light of a non-being …' *Being and Nothingness* (Sartre 1984: 563). Later, focus shifted to freedom as a function of meeting material needs of human beings; 'It would be quite wrong to interpret me as saying that man is free in all situations … . I mean the exact opposite: all men are slaves in so far as their life unfolds … always conditioned by scarcity' *Critique of Dialectical Reason Volume One* (Sartre 2004: 332). This was a huge shift, and I will endeavour to outline features of each perspective.

The essential starting point is the claim that freedom for human beings *is* the very nature of human consciousness (being-for-itself). Consciousness has no inherent content, it is in an important sense nothing, but can define *itself* at any moment. It is future orientated and seeks content: 'This is its freedom. Freedom allows the for-itself to redefine itself in every instant, it gives it the power to break from the past and to redefine the future' (Franchi 2020). On this view, normal conscious human beings are condemned to be free (Sartre 1984: 567), unlike inert matter it is the very nature of their being, hence the description 'ontological freedom'. It will be quickly realised that the claim of ontological freedom has counter intuitive implications, in that freedom is not considered contingent; a prisoner is free because choices are available concerning how to respond to being in prison (Sartre 1984: 622).

The reasons for transition to essentially a material conception of freedom are the subject of much discussion, but it is generally believed that the Second World War and the Holocaust were instrumental in bringing about this change. The idea that human beings are free in any situation was surely unsustainable given the nature of what happened during the war years, as expressed for example in excerpts from *Anti-Semite and Jew* (Sartre 1976) published in *Temps Modernes* in 1945. The notion of free in chains (Sartre 1984: 703) was not entirely discarded as an ontological quality of being human but had been shown as a grossly incomplete characterisation of human freedom. What was lacking was a material conception of freedom.

Consideration of material freedom begins with choice: As suggested above, the ontologically free prisoner detained in Alcatraz has a choice, to accept captivity or try to escape by swimming over a mile from Alcatraz Island to the San Francisco shoreline, but these are clearly *poor* choices. Choice does not necessarily mean freedom, in a normal and plausible sense. Being oppressed or imprisoned is not all about absence of choice but being forced to choose between bad options. Material freedom requires freedom from coercion and domination. Note the radical change from an ontological based notion of freedom to considerations that are independent of human nature; for example, considerations based on human rights and material security, availability of food, water, shelter and education, and freedom from slavery, poverty, discrimination, domination and persecution.

Material freedom for Sartre is essentially basic material security, an absence of coercion, but has been articulated in terms of access to cultural and social goods necessary for pursuing personal projects and ambitions, (following The Internet Encyclopedia of Philosophy *Sartre's Political Philosophy* Section 3: Freedom, based on Sartre's *Notebooks for an Ethics* (1992: 329-331)).

The above offers a very brief sketch of Sartre's notions of human freedom; description of Sartre's ontological perspective is slight, but unfortunately giving justice to the complexity and scope of Sartre's philosophy on this point as part of a very much larger and complex interlocking system of ideas is unfortunately beyond the scope of this Thesis.[46]

A much wider context of Sartre and Existentialism may be found in *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience* (Flanagan and Caruso 2018). Three 'waves' of existentialism are described. The first, relates to anxiety about justification of moral norms without theological foundation. The second, anxiety over the catastrophic failure of human goodness and rationality to respond to this absence, as manifest, for example, by the Second World War, Stalin and Pol Pot as mentioned above. Third, Neuroexistentialism; anxiety caused by the rise of a 'scientific image' implying denial of free will and any ultimate human outcome other than death, as per all other mammals (following Flanagan and Caruso 2018: 2).[47]

1.2 Why Free Will is an important issue

There are numerous expressions of the 'free will problem' and the importance of the notion of free will generally, in the context of moral responsibility, the law, our concept

[46] In this section I follow closely Storm Heter's article from The Internet Encyclopedia of Philosophy *Sartre's Political Philosophy* Section 3: Freedom <https://www.iep.utm.edu/sartre-p/#H3>. For a detailed and clear exposition of Sartre's ontological perspective on freedom see Gary Cox, *Sartre: Consciousness, Freedom, Bad Faith* (2014). Alternatively, Diané Collinson and Kathryn Plant *Fifty Major Philosophers* (2006) presents an exceptionally clear description of Sartre's concept of consciousness, described by the authors as 'perhaps a difficult idea to grasp' (2006: 231).

[47] This is a *gross* simplification of Flanagan and Caruso's wide ranging and hugely interesting Paper.

of ourselves as persons, and so on. This is one such concise and eloquent expression in the context of consciousness:

> For any materialist vision of consciousness, the crucial stumbling block is the question of free will. A modern, enlightened person tends to feel that he or she has rejected a mystical, immaterial conception of the eternal soul in exchange for a strictly scientific understanding of consciousness and selfhood – as something created by the billions of neurons in our brains with their trillions of synapses and complex chemical and electrical processes. But the fact of our being entirely material, hence subject to the laws of cause and effect, introduces the concern that our lives might be altogether determined. Is it possible that our experience of decision-making – the impression we have of making choices, indeed of having choices to make, sometimes hard ones – is entirely illusory? Is it possible that a chain of physical events in our bodies and brains must cause us to act in the way we do, whatever our experience of the process might be? (Parks and Manzotti 2020)

I believe the brief historical perspective of Section 1.1 clearly and concisely reflects, confirms and demonstrates the 'problem' of human freedom; to understand 'the extent to which human agents *are* in control of their own lives and destinies' (added emphasis Russell 2013b: 2) when threatened, for example, by fate, theological issues or physical determinism. Seeking to understand the tension between the internal sense of freedom experienced by most human beings and freedom regulating, perhaps eliminating, features of the world and our religious beliefs is an ongoing search across many cultures and centuries that is self-evidently important. Important in itself, *and* because of many vital consequences that follow from our understanding of the extent or possibility of human freedom, such as, responsibility for actions, giving praise, blaming and punishment.

1.3 Summary

Chapter 1 offers reflections on free will based on the work of some major philosophers and theologians. The importance of the free will debate is confirmed across a multitude of areas, for example, moral responsibility, metaphysics and jurisprudence. I have introduced the free will debate that will continue over the next two chapters, showing semicompatibilism developing from a tradition of thought concerning the control that human beings have over their lives that began thousands of years ago. This provides a foundation and context for Part III where the impact of implicit bias on a particular free will position will be examined. The intention is to increase focus through the Chapters of Part I, concluding with a description of semicompatibilism in Chapter 3.

What is the problem? With Nagel (2003), I believe 'the essential problem' can be summarised as the challenge of resolving the tension between our self-image as agents in control of our lives, making rational, responsible decisions in a way that makes us feel like things are, in an important sense, 'up to us' and, as outlined, the various challenges to such feelings of agency and so responsibility.[48]

How a particular understanding or concept of freedom and responsibility (semicompatibilism) responds to the challenge of implicit bias is the essential issue to be addressed within Part III.

[48] Such 'challenges' to our agency may also be unconscious, for example, some understandings of implicit bias.

Chapter 2

Contemporary Responses to the Free Will Problem

We know our will is free, and there's an end on't.
Samuel Johnson[49]

2.0 Introduction

Introducing Part I, Chapter 1 outlined the ideas of some important philosophers and writers on human freedom. Reflection on freedom essentially seeks to answer two questions; is free will possible, and if it is, what is its nature? The possibility of free will and freedom of action is usually considered together with, or despite of, forms of determinism such as those described, the prophecies of the Oracles of classical antiquity, physical causation or psychological determining factors. Critical examination attempts to understand if and under what circumstances there could be compatibility between apparently inconsistent concepts of freedom and forms of determination. In this chapter I will outline responses to the free will problem expressed in terms of claims and arguments about compatibility or incompatibility of freedom and material determinism (as outlined, for example, by Parks and Manzotti (2020) on page 32), and implications for the vital notion of moral responsibility. I will briefly describe the main positions and arguments concerning compatibility of freedom and determinism, (the possibility and nature of free will), before concentrating in Chapter 3 on the semicompatibilist position. Recall, the phenomenon of implicit bias challenges the semicompatibilist position if issuing behaviour is found to be immune from the type of control needed for responsibility (guidance control) in the presence of substantial evidence and argument supporting the claim that agents *are* responsible. Investigating the nature of this possible threat is the essential research activity within this Thesis.

[49] *The Life of Samuel Johnson* (Boswell 1998).

Before looking at compatibility issues and various free will positions in more detail, I will mention three basic concepts employed extensively throughout the free will debate. The Introduction to *Four Views on Free Will* claims 'perhaps the three most important concepts in philosophical work on free will are free will, moral responsibility, and determinism' (Fischer et al. 2007:1). First, Robert Kane in the Introduction to *The Oxford Handbook of Free Will* describes our sense of free will when

> we think of ourselves as capable of influencing the world in various ways. Open alternatives seem to lie before us. We reason or deliberate among them and choose. We feel it is 'up to us' what we choose and how we act; and this means that we could have chosen or acted otherwise … the origins or sources of our actions lie in us and not in something else over which we have no control - whether that something else is fate or God, the laws of nature, birth or upbringing, or other humans. (2012: 4)

Kane's quotation covers the key points concerning what it feels like from the inside to have freedom of will and action.[50] However, there is unfortunately no single and agreed understanding of free will or the free will problem. There are

> […] a range of problems … a troubling entanglement of our concepts, an entanglement that seems to lead to contradictions … to settle the problem - to disentangle the set - we must either reject some concepts, or instead, we must demonstrate that the set is indeed consistent despite its appearance to the contrary. (McKenna and Coates 2015a)

The concepts that *are* rejected or shown to be (in)consistent, despite appearances to the contrary, define positions within the free will debate. This is important and may be developed using what is often referred to as the Classical Formulation of the free will problem (McKenna and Coates 2015a):

1. Some agents, at some time, could have acted otherwise than she did.
2. Actions are events.
3. Every event has a cause.
4. If an event is caused, then it is causally determined.

[50] See also Thomas Nagel *Freedom* (2003) for further description of our sense of freedom from the inside.

5. If an event is an act that is causally determined, then the agent of the act could not have acted otherwise than in the way that she did.

From the Classical Formulation distinct positions within the free will debate emerge. There are six concepts; an agent, action, could have done otherwise, event, cause and causal determination. There is also a mutually inconsistent set of propositions, yet each one seems rooted in our contemporary conception of the world: Proposition (1) is grounded in a conception of agency (free agency) as an ability to select among different possible courses of action. This idea is sometimes described by analogy between alternative possibilities and a Garden of Forking Paths, whereby an agent's future is the result of ongoing choices between alternative forking paths that branch off from a single past; from a single past, there is for the agent more than one possible path into the future. [51] Proposition (2) identifies actions with events. Proposition (3) is a presupposition of natural science. Indeterminacy and uncaused events will be mentioned later. Proposition (4) has historically been a working assumption of the natural sciences. Proposition (5) arises from a common-sense understanding of what it means to claim that an event is causally determined - given the antecedent causal conditions for the event, it was not possible for it not to have occurred (following closely McKenna and Coates 2015a). With respect to this formulation, compatibilists deny (5), but incompatibilists claim free will and determinism are incompatible and support (5). Note, there is no commitment yet as to the *truth* of determinism or alternatively that anyone has free will, what McKenna calls an agnostic incompatibilism. An incompatibilist hard-determinist thesis does commit to the claim that no person has free will and determinism *is* true. Finally, the incompatibilist libertarian accepts the truth of free will *and* denies the truth of determinism. In terms of the Classical Formulation, libertarians may deny (3) claiming that causation is indeterministic or deny (4) claiming if events *are* caused then such causation is not deterministic. Clearly (1) is fundamental and controversial, a central issue within virtually any discussion of free will.

Responsibility, the second basic concept, is linked to others such as desert, accountability, blameworthiness and praise. The term responsibility may be used in the

[51] The origin of this term is *The Garden of Forking Paths,* a 1941 short story by Argentine writer and poet Jorge Luis Borges.

sense of an obligation, but it is the sense of responsibility and its connection with praise and blame that is important here (Fischer et al. 2007: 2). As McKenna points out:

> A person who is a morally responsible agent is not merely a person who is able to do moral right or wrong … she is *accountable* for her morally significant conduct. Hence, she is, when fitting, an apt target of moral praise or blame, as well as reward or punishment. (added emphasis McKenna and Coates 2015a)

The claim that for an agent to *be* morally responsible it is necessary to have the ability to make free decisions is widely held and intuitively plausible, but it is far from universally agreed. Some philosophers express the connection between responsibility and freedom from a different point of view, using the ability to take genuinely responsible decisions and actions as a benchmark or defining factor *for* freedom. For example, McKenna offers a definition of free will as 'the unique ability of persons to exercise control over their conduct in the manner necessary for moral responsibility' (McKenna and Coates 2015a). When free will is considered in this way, if responsibility could legitimately be present in the absence of alternative possibilities, then perhaps free will (at least in some important sense, for example, semicompatibilism) could also exist without alternatives, (in this 'strong' sense), available to the agent; a situation that may initially seem implausible. However, arguments have been developed, (the most well-known and perhaps most important by Harry Frankfurt), that claim an agent may be responsible for their actions even in the absence of alternative possibilities. Arguments of this type are now generally referred to as Frankfurt-type examples or Frankfurt-style examples and will be discussed shortly. Frankfurt-type examples will be seen to be very important as a major motivator of the semicompatibilist free will position.

The third basic concept, determinism, is essentially the thesis that at any time the universe has exactly one physically possible future. Something is deterministic if it has one physically possible outcome given a particular set of starting conditions and fixed causal laws, usually referred to as 'the laws of nature' (following Fischer et al. 2007: 2). The well-known argument for incompatibility, the consequence argument, can help to clarify the general notion of determinism, see particularly clause 4:

1. There is nothing we can now do to change the past.
2. There is nothing we can now do to change the laws of nature.
3. There is nothing we can now do to change the past and the laws of nature.
4. *If* determinism is true, our present actions are the necessary consequences of the past and the laws of nature. (That is, it must be the case that, given the past and the laws of nature, our present actions occur).
5. Therefore, there is nothing we can now do to change the fact that our present actions occur.

Forms of determinism threaten free will by taking away the key element of 'up to us' when apparently (from the agent's perspective) choosing from alternative possibilities since only one alternative is actually possible. The Classical Incompatibilist Argument below (McKenna and Coates 2015a), shows, if determinism is true, there is no access to alternatives as illustrated by the Garden of Forking Paths:

1. If a person acts of her own free will, then she could have done otherwise.
2. If determinism is true, no one can do otherwise than one does.
3. Therefore, if determinism is true, no one acts of her own free will.

An alternative argument that shows the apparently controlling or freedom denying nature of determinism is expressed in terms of the idea that for an action to be free it must originate entirely from within the agent. McKenna (2015a) calls this the Source Incompatibilist Argument:

1. A person acts of her own free will only if she is its ultimate source.
2. If determinism is true, no one is the ultimate source of her actions.
3. Therefore, if determinism is true, no one acts of her own free will.

The idea of 'ultimate source' is interesting and important. Proposition (1) seems initially reasonable, fitting well with intuitive notions of free will, but how is the ultimate source 'she' to be understood? Reflection quickly identifies how difficult it is to describe, understand or imagine an aspect of our nature that could exist independent of external causation or influence and yet *be* causal within the world.[52] Perhaps such difficulties are resolved if 'we really are souls, immaterial and immortal clumps of God stuff that inhabit and control our material bodies … ' (Dennett 2003: 1), but this is scarcely an uncontroversial claim. The notion of agent causation is a fundamental issue within the free will debate and will be discussed again. The main positions within the free will debate, distilled into a few words, are shown in Fig 2.1 (Fischer et al. 2007: 4). See also Daniel Dennett and Gregg Caruso's clear and concise comparison of the main contemporary free will positions in the Introduction to *Just Deserts* (2021: 2).

	Is common-sense thinking about free will and moral responsibility basically correct?	Is free will compatible with determinism?	Is moral responsibility compatible with determinism?	Do we have free will?
Libertarianism	Yes	No	No	Yes
Compatibilism	Yes	Yes, although a semicompatibilist may say 'no'[53]	Yes	Yes
Hard Incompatibilism	No	No	No	No
Revisionism	No	Yes, but only with revision to our self-image	Yes	Yes, or close enough

Fig 2.1 An overview of the main responses to the free will problem.

[52] See description and discussion of emergence and agent causation in Appendix B.

[53] This may appear confusing; semicompatibilism 'puts to one side' the question of the truth of determinism. Even if determinism is true, an agent *has* freedom/responsibility, *but* freedom/responsibility as defined within semicompatibilism, not necessarily freedom in the 'Forking Paths' sense, see Chapter 3 Semicompatibilism, for a more detailed explanation.

2.1 Libertarianism

Incompatibilists, as the name suggests, believe free will and determinism are incompatible. Libertarian incompatibilists claim that we *do* have free will and thereby claim determinism to be false. In other words, incompatibilists

> believe free will and determinism are incompatible. If incompatibilists also believe that an incompatibilist free will exists, so that determinism is false they are called libertarians about free will. (Fischer et al. 2007: 7)

However, if our actions are not determined, at least in some sense, the alternative world of undetermined actions seems a strange and unrecognisable place where notions of free will and responsibility continue to challenge understanding, because undetermined actions seem random and spontaneous, and therefore are not responsible actions. If libertarians are correct when claiming free will exists and is incompatible with determinism then what is the source, if any, of actions and how is responsibility to be understood and justified? As Randolph Clarke and Justin Capes comment in their introduction to *Incompatibilist (Nondeterministic) Theories of Free Will* 'the task of providing an incompatibilist account is not an easy one' (Clarke and Capes 2017).

To give an account of *any* position that tries to show compatibility *or* incompatibility between determinism and free will is not easy. However, Robert Kane does offer a detailed and persuasive incompatibilist libertarian argument. Kane argues that any defence of the libertarian position must show (i) free will really is incompatible with determinism, and (ii) libertarian free will requiring *indeterminism* can be intelligible and compatible with current scientific knowledge. In other words, both elements of the libertarian position must be covered, incompatibility *and* the existence of free will. Kane discusses the previously mentioned consequence argument (Fischer et al. 2007: 10):

1. There is nothing we can now do to change the past.
2. There is nothing we can now do to change the laws of nature.
3. There is nothing we can now do to change the past and the laws of nature.
4. *If* determinism is true, our present actions are the necessary consequences of the past and the laws of nature. (That is, it must be the case that, given the past and the laws of nature, our present actions occur).

5. Therefore, there is nothing we can now do to change the fact that our present actions occur.

After a detailed and careful analysis Kane shows potential problems with this argument. The past and laws of nature cannot be changed, (surprisingly, a claim not universally accepted), but does such powerlessness *thereby* transfer to actions that are willed and are within our power, (actions that are not logically or physically impossible), to perform today? Arguments are ongoing about the legitimacy of the consequence argument, but Kane presents an alternative position based on 'ultimate responsibility'. For an agent to be *ultimately* responsible *they* must be responsible for anything that is a sufficient cause or motive for the action to occur; cause or motive that may *originate* from within their own character:

> If … an agent's choice issues from, and can be sufficiently explained by, an agent's character and motives (together with background conditions), then to be ultimately responsible for the choice, the agent *must* be at least in part responsible by virtue of choices or actions voluntarily performed *in the past* for having the character and motives he or she now has. (added emphasis Fischer et al. 2007: 14)

The vital point is that ultimate responsibility does not entail that we could have done otherwise for *every* act that is performed. However, such ultimate responsibility *must* entail an undetermined choice that is incompatible with determinism, a freedom worth wanting with respect to *some* acts in our past that form our present character. Kane calls these earlier acts by which we formed our present character 'self-forming actions' (Fischer et al. 2007: 14). Brief reflection on the idea of 'self-forming actions' and Kane's related notion of 'will-setting actions' suggests a clear regress problem; self-forming actions and will-setting actions occurring in the context of earlier self-forming actions regressing to infancy. At some point, for ultimate responsibility to make sense, there must be at least one undetermined choice. How can the idea of an undetermined choice be understood, as to be undetermined is surely to be arbitrary and so the entire process of character formation begins grounded on luck? Kane claims otherwise, saying it is "a mistake, (in fact, one of the most common in debates about free will), to assume that 'undetermined' means 'uncaused' or 'merely a matter of chance' "(Fischer et al. 2007:

31). Kane argues that the effort of the agent and the important element of indeterminism are fused, not something separated in time. It is the effort of the *agent*; hence the outcome may quite legitimately be said to be the agent's responsibility. The role of indeterminism is seen as a hindrance or resistance[54] that 'paradoxically opens up the genuine possibility of pursuing other purposes – of choosing or doing otherwise in accordance with, rather than against, our wills (voluntarily) and reasons (rationally)' (Fischer et al. 2007: 39). Concerned that lingering doubts remain about indeterminism having a legitimate role in free will, Kane agrees that self-forming actions do have an element of arbitrariness but believes this shows something important about free will. Every self-forming free choice is in some sense an experiment that looks to the future for justification and cannot be explained fully in terms of past events.

There are, of course, many and varied arguments supporting, developing and raising objections against libertarianism. Just one supporting argument has been sketched here, but one that is influential and developed by Robert Kane, probably the leading libertarian within the current free will debate.

2.2 Compatibilism

Many philosophers argue that even though it may initially appear impossible, determinism does *not* in fact threaten free will; certainly any form of free will 'worth wanting' (Dennett 1984). As described, scholarship relating to free will generally and compatibilism in particular is vast, therefore summarising such a variety of positions is challenging, but Michael McKenna's historical overview is a good starting point:

> A useful manner of thinking about compatibilism's place in contemporary philosophy is in terms of at least three stages. The first stage involves the classical form defended in the modern era by the empiricists Hobbes and Hume and reinvigorated in the early part of the twentieth century. The second stage involves three distinct contributions in the 1960's, contributions that challenged many of the dialectical presuppositions driving classical compatibilism. The third stage

[54] Kane uses an image from Kant's *Critique of Pure Reason* to help explain the idea of indeterminism as a positive resistance or hindrance; the image of a bird that is 'upset by the resistance of the air and the wind to its flight and so imagines that it could fly better if there were no air at all to resist it. But of course, the bird would not fly better if there were no air. It would cease to fly at all. So, it is with indeterminism with free will. It provides resistance to our choices, but a resistance that is necessary if we are to be capable of true self-formation' (Fischer et al. 2007: 40).

> involves various contemporary forms of compatibilism, forms that diverge from the classical variety and that emerged out of, or resonate with, at least one of the three contributions found in the second transitional stage. (McKenna and Coates 2015a)

The First Stage - Classical Compatibilism. Classical compatibilism essentially claims that to be free is to have the power or ability to do what one wills to do, free of constraints that would prevent acting following what is willed. If determinism is true, it does not entail that the will or action is impeded or compelled in a *relevant* way; free will can be simply expressed as the *unencumbered* ability to do what one wants.[55] If determinism is true, no encumbrance to action is thereby created. If the classical compatibilist's wants and actions are determined yet physically unencumbered then the agent, based on an unrefined account of determinism, cannot actually do otherwise; while the classical compatibilist's condition for free will may be satisfied, the impossibility to act otherwise runs counter to currently held notions of what it is to act freely.

The Second Stage – Three Major Contributions. Three major contributions during the 1960's radically changed the free will debate. McKenna claims 'No account of free will, compatibilist or incompatibilist, is advanced today without taking into account at least one (if not more) of these three pieces' (2015a: Section 4):
(i) The incompatibilist argument developed by Carl Ginet, known as the Consequence Argument, described on page 37.
(ii) Harry Frankfurt's argument, contra the Principle of Alternate Possibilities (PAP), that an agent unable to do otherwise may nevertheless be morally responsible. While looking at responsibility, I noted that if it could be shown that responsibility could legitimately be present in the absence of alternative possibilities, then perhaps a substantial, meaningful, sufficient human freedom, (such as expressed by semicompatibilism), could also exist without alternative paths being available to the agent. I will describe how Harry Frankfurt pushed back against the Principle of Alternate Possibilities, a central idea in the free will debate and a principle commonly employed in the wider community. Appearing in *The Journal of Philosophy*, under the title 'Alternate Possibilities and Moral

[55] For detailed discussion of the related subject of duress see Carla Bagnoli, *Claiming Responsibility for Action Under Duress* (2018).

Responsibility' (1969), Harry Frankfurt's hugely important argument, or thought experiment, counters the plausible claim that an agent is morally responsible for what they have done *only* if they could have done otherwise (PAP). Frankfurt's argument typically proceeds as follows: Assume that a controller has power over an agent's actions, perhaps by direct control of the agent's brain. The controller will not intervene if the agent is going to do on her own what the controller wants. Frankfurt argues that if the controller does not intervene because the agent performs the desired action entirely on her own, the agent can be morally responsible for what she does (since the agent acted on her own and the controller was not involved), even though the agent literally could not have done otherwise, because the controller would intervene and not allow it (following Kane 2012). So, the Principle of Alternate Possibilities seems to have been countered with respect to responsibility.[56] To be clear, Frankfurt style examples describe an agent who is morally responsible even though it is impossible for the agent to do otherwise than they actually did. From the agent's point of view, actions are chosen freely even though, unknown to the agent, there are no actual alternative possible actions. The controller of the Frankfurt example is always present but passive while the agent is doing what the controller wants and active only if the agent begins to stray off course.[57] The agent's position bears some similarity to the sleeping man, in John Locke *An Essay Concerning Human Understanding*, (Book II, Chapter xxi, item 10), unaware that his room has been locked from the outside during the night. In the morning he decides to remain in the room, still unaware that his room is locked. Although he cannot leave the room, unaware of his true condition, he believes (he feels) that he is free to remain or leave. Obviously, Lock's sleeping man will realise his true situation when trying to leave, whereas Frankfurt's agent presumably remains unaware of the Controller's presence or intervention.

A great deal of energy and ink has been devoted to analysis of Frankfurt-type examples, a trend that will continue to some small extent within this Thesis. Clearly, the

[56] Frankfurt offers an alternative version of PAP that considers objections raised earlier within *Alternate Possibilities and Moral Responsibility:* 'a person is not morally responsible for what he has done if he did it only because he could not have done otherwise' (1969: 838).

[57] For detailed discussion of possible meanings of 'begins to stray off course' see *Living Without Free Will* (Pereboom 2003: 28-33).

role of Frankfurt's Controller cannot simply be read across to the role of implicit bias, but there is perhaps some small similarity between implicit bias and the Controller as an unknown behaviour determining factor that becomes active in particular circumstances.[58]

(iii) P. F. Strawson's important and influential defence of compatibilism invites both compatibilists and incompatibilists to attend more carefully to the central role of interpersonal relationships and the reactive attitudes in understanding the concept of moral responsibility[59] (following McKenna and Coates 2015a).

The Third Stage - Contemporary Forms of Compatibilism. Before looking at some specific contemporary forms of compatibilism, (sections 2.2.1 to 2.2.5), after a brief introduction I consider the opposing incompatibilist position in more detail and then give a brief overview of how compatibilists have responded to some difficult challenges. I will also describe an important challenge to compatibilism, usually referred to as the manipulation argument.

The compatibilist position is a broad church. Quoting Galen Strawson, Pereboom illustrates the diversity of positions within compatibilism, where supporters can believe any of at least the following (Pereboom 2003: xvi):

- That determinism (D) is true, that D does not imply that we lack the free will required for moral responsibility (F), that we in fact lack F.
- That D is true, that D does not imply that we lack F, that it has not been shown whether (or not) we have F.
- That D is true, and that we have F.
- That D is true, that we have F, and that our having F requires that D be true.
- That D may or may not be true (i.e., we do not know whether D is true), but in any case, we have F.
- That D is not true, that we have F, and would have F even if D were true.

[58] Recall earlier description of freedom within the limitations imposed by God's foreknowledge and situations where agents act freely in the sense that their actions are what *they* want to do but are nonetheless ultimately in accord with God's will, Oracles and so on.

[59] See P. F. Strawson's seminal Paper, *Freedom and Resentment* (1962).

- That D is not true, that we do not have F, that F is nonetheless compatible with D.

It is worth noting that confusion may arise concerning the terms compatibilist and compatibilism. Compatibilism can be considered as a purely metaphysical problem concerning the compatibility of free will and determinism independent from any view or argument about moral responsibility and moral obligation. From this perspective the *outcome* of the free will – determinism problem has obvious relevance for moral philosophy, but the compatibility problem is positioned strictly within metaphysics. By contrast, 'compatibilist' is often used in ways that *only* include the moral issue of responsibility and determinism; essentially, it is important to be clear about what is included within any compatibility claim and what is excluded. Note how Galen Strawson defines (F) in the first item above.

It was noted on page 40 that the compatibilist's essential task of defending some form of compatibility between free will and determinism is not easy. By contrast, describing *challenges* faced by compatibilists can be relatively straight forward. For example, from the perspective of freedom and physical determinism, typically a challenge can be expressed as follows.[60] If human beings inhabit the world and consist of common materials arranged in complex ways that form our physical structure, then it appears reasonable and plausible to claim that what is true of the rest of the world is also true of human beings. Given the world is characterised by a multitude of physical/mechanical causal relationships that are predictable and consistent, allowing science and technology to flourish, it seems rational to claim that human beings also follow such predictable and consistent behaviour, acting in accordance with the laws of nature. If this is not the case, and we are not held within a chain of cause and effect, then something unique, something specific to humans, (and probably many animals), must be present or some additional consideration available, (such as the notion of agent causation), to explain how humans are fundamentally different from everything else that we currently know exists, by virtue of exemption from the constraints of general

[60] Thomas Nagel expresses this point eloquently at the beginning of his article Freedom, *Oxford Readings in Philosophy; Free Will* (2003: 229).

mechanical/physical causation.[61] We *feel* we act freely,[62] that freedom is an essential part of what it is to be human but if this is true it raises obvious questions regarding how such freedom is possible. In a world of pervasive determinism how can a fragment or flicker of freedom be found, enabling freedom and/or moral responsibility to exist?

Recall, the issue may be expressed succinctly in terms of incompatibility between propositions of the Classical Formulation (McKenna and Coates 2015a):

1. Some agents, at some time, *could* have acted otherwise than she did.
2. Actions are events.
3. Every event has a cause.
4. If an event is caused, then it is causally determined.
5. If an event is an act that is causally determined, then the agent of the act *could not* have acted otherwise than in the way that she did.

From earlier discussion, this is familiar territory; to respond to this apparently inconsistent, yet individually plausible, set of propositions the compatibilist must develop a position that counters or responds appropriately to proposition 5. For example, by cultivating a more nuanced understanding of determinism and/or acting otherwise. If the above set of propositions are simplified (McKenna and Coates 2015a) the basic incompatibilist challenge becomes very clear:

1. If a person acts of her own free will, then she could have done otherwise.
2. If determinism is true, no one can do otherwise than one actually does.
3. Therefore, if determinism is true, no one acts of her own free will.[63]

[61] Susan Wolf expresses a similar point, (more succinctly), see *Freedom Within Reason* (1993: 70).

[62] For information concerning the construction and validation of a psychometric tool for measuring beliefs about free will and related concepts see *The Free Will Inventory: Measuring Beliefs about Agency and Responsibility* (Nadelhoffer et al. 2014).

[63] There are two fallacies to keep in mind when talking about free will and responsibility: First, the ethical fallacy, that free decisions *must* be moral decisions. While it is arguably true that free decisions are necessary to be held morally responsible, they do not of themselves necessarily lead to moral behaviour. Second, the rational fallacy, that a free choice must be a rational choice. Further, as described in Chapter 1 determinism takes different forms and sight must not be lost of determinism's wider perspective beyond physical determinism. Additional information regarding the ethical and rational fallacy may be found at;

Having described some difficulties raised against the compatibilist position by incompatibilist arguments, how does the compatibilist respond? Michael McKenna and Justin Coates offer a working definition of free will as 'the unique ability of persons to exercise *control* over their conduct in the fullest manner necessary for *moral responsibility*' (added emphasis 2015a). (Recall my important comment (page 37) concerning inclusion of moral responsibility within the definition of free will). This leads to two possible facets of control.

The first facet is control in the sense of being able to select *sufficiently* freely between alternative courses of action. While this form of control is often considered to be *regulative* control, following John Martin Fischer (2013), it will be seen that sufficient control may *not* have to be as 'strong' as regulative control to bring about agent responsibility. The compatibilist must give a convincing account of agent control concerning choice between alternative actions that is substantial enough to satisfy, for example, McKenna and Coates' (and many others) requirement of moral responsibility *for* that choice. John Martin Fischer develops such a model of control, (*guidance* control), within his free will semicompatibilist position.

An incompatibilist challenge may also be made based on the consequence argument, a very compelling and difficult argument for the compatibilist to counter. Recall, that assuming determinism is true:

1. No one has power over the facts of the past and the laws of nature.
2. No one has power over the fact that the facts of the past and the laws of nature entail *every fact* of the future (i.e., determinism is true).
3. Therefore, no one has power over the facts of the future.

Pushing back against the consequence argument appears to be particularly difficult if the chosen target is premise 1. Surely it is the case that this premise is impervious to counter argument? However, subtle discussion and interesting arguments[64] have been developed

The Information Philosopher <http://www.informationphilosopher.com/freedom/ethical_fallacy.html> and <http://www.informationphilosopher.com/freedom/rational_fallacy.html>.

[64] See for example, John Turk Saunders, *The Temptations of 'Powerlessness'* (1968), Wesley Holliday, *Freedom and the Fixity of the Past* (2012) and David Lewis, *Are We Free to Break the Laws?* (1981).

that challenge or at least raise questions concerning this premise. For example, by developing the 'difference between a person in the present who has the ability to act in such a way that *she alters the past*, as opposed to a person who has the ability to act in such a way that, *if she did so act, the past would have been different'* (McKenna and Coates 2015a). It is beyond the scope of this Thesis to critique these very nuanced and difficult challenges to the fixity of the past and laws of nature, but it should be noted that even such apparently uncontroversial claims as premise 1 continue to be subject to intense scrutiny.

Another possibility for the compatibilist to push back is based on the claim that while premise 1 and 2 concern matters that have nothing to do with a person's agency, premise 3 *does* relate to individual agency. It has been argued that the *inference* from the unavoidability within premises 1 and 2 to the unavoidability of actions by the agent within conclusion 3 is incorrect, (see John Turk Saunders, *The Temptations of 'Powerlessness'* (1968), for detailed discussion of fallacies, mistakes and confusion concerning proofs of powerlessness. Also, Patrick Grim *Free Will in Context: A Contemporary Philosophical Perspective*, Sections 3 and 4 (2007) for discussion of inference mistakes in arguments of this form). The term 'every fact' within premise 2 seems to counter this claim, however, it is acknowledged that such a brief look does not provide a sufficiently full or fair account of these fine-grained responses to the consequence argument. It is worth noting that even if a compatibilist were to successfully refute the consequence argument, or any incompatibilist argument, a full explanation requires a *positive* account of, or response to, determinism and control such as semicompatibilism, developed by John Martin Fischer in *Four Views on Free Will* (2007: 44), *My Compatibilism* (2013: 296-317) and particularly, *Deep Control* (2015).

The second facet is control in the sense of the agent's self or real/deep self as the ultimate source of action or behaviour. This second facet, source control, requires the compatibilist to describe the nature of a 'real self' that is sufficiently causally isolated from the determined world and yet able to be the agent's ultimate source *of* causation and action within the world. So, the second facet, requiring agent isolation in some form or degree from pervasive determinism (yet causally active in the world), *and* the presence within the agent of a source of behaviour that is truly 'their own' is considered sufficient to ensure responsible behaviour. This topic is discussed in greater detail in Part II in the context of control of implicit bias related behaviour, also in Appendix B Agent Causation.

Responding to some incompatibilist challenges, when free will is defined as 'the unique ability of persons to exercise control over their conduct in the fullest manner necessary for moral responsibility' (McKenna and Coates 2015a), the two forms of control mentioned above, (guidance control and source control), will be discussed again in much greater detail in later chapters.

Concluding this brief description of some of the major challenges for the aspiring compatibilist, I will describe a form of argument known as the manipulation argument. Manipulation arguments are generally of the following form: an *in*compatibilist presents a state of affairs that (it is claimed) satisfies stated sufficient compatibilist conditions for moral responsibility (and freedom), but crucially, the incompatibilist then describes how this exact state of affairs *could* have been brought about through some form of agent manipulation. Being manipulated in this way suggests the agent is in fact *not* responsible, therefore such compatibilist conditions are (at least) insufficient for moral responsibility (and freedom). In other words, an incompatibilist manipulation argument (following McKenna and Coates 2015b) maintains that an agent so manipulated is *not* free or morally responsible despite satisfying pertinent compatibilist-friendly conditions *of* responsibility and freedom. Further, any agent who is determined to perform X is not different in any relevant respect from an agent manipulated into performing X. Therefore, when compatibilist's conditions are satisfied in cases of manipulation they could equally well be satisfied under conditions where determinism is true; therefore, such conditions are not sufficient to guarantee *independence* from possible determinism with associated freedom and responsibility. More succinctly, ' … regarding moral responsibility, there is no important difference between various cases of manipulation in which agents who A are not morally responsible for A-ing and ordinary cases of A-ing in deterministic worlds'[65] (Mele 2005: 75). As expected, running together manipulation

[65] Perhaps the most well-known expression of the manipulation argument is presented by Derk Pereboom, often referred to as the Four Case Argument (2003: 112). This argument has been described as 'a kind of litmus test for the credibility of a compatibilist theory', concluding that the 'source of the agent's actions can be traced back in their entirety to originating conditions that were completely beyond her control' (McKenna and Coates 2015b). For a critique see Alfred R. Mele *A Critique of Pereboom's 'Four-*

and determinism as described has been challenged, essentially by claiming a distinction between modes of causation that undercut responsibility and freedom and others that do not. In the next chapter it will be seen that semicompatibilists argue their comprehensive conditions for guidance control and so responsibility are *not* satisfied in cases of manipulation.

Having described challenges to the compatibilist position and the compatibilist response, I will continue with the third stage of Michael McKenna's (2015a) historical overview, discussion of some contemporary forms of compatibilism, including the freedom to do otherwise, hierarchical compatibilism, the reason view, reasons-responsive compatibilism and Strawsonian compatibilism.

2.2.1 Freedom to do Otherwise - Dispositionalism

Positive compatibilist accounts of determinism and regulative control, in the sense of being able to select freely between alternative courses of action, have been developed that respond to premise two of the classical formulation; if determinism is true, no one can do otherwise than one actually does. The difficulty here is plain to see, given the uncompromising requirement of such regulative control. Dispositionalism is an attempt to offer what appears to be impossible, a positive account of 'can do otherwise' in a world where determinism may be true. An account of Dispositionalism is given by Kadri Vihvelin in her clear and eloquent book *Causes, Laws, and Free Will: Why Determinism Doesn't Matter* (2013: 171). It is claimed that 'we have the free will we think we have by having some bundle of narrow abilities *and* by being in suitably friendly surroundings; when this is so we have not only the narrow but also the wide ability to do otherwise' (2013: 167). Narrow ability is intrinsic and shared by most human beings, enabling decisions to be made based on reasons and having the intrinsic power or agency to act otherwise, based on such decisions. Narrow abilities are like intrinsic dispositions of objects; dispositions such as fragility, elasticity, solubility, and so on. Wide abilities are those abilities by virtue of having narrow abilities, together with further facts about surroundings. It is relatively easy to change someone's wide abilities by changing their surroundings, (such as locking a door), but removing someone's narrow abilities requires

Case Argument' for Incompatibilism (2005) also Mele's vital Paper *Manipulation, Compatibilism, and Moral Responsibility* (2008).

an intrinsic change be made to the person; even if determinism is true there is no reason to believe that narrow abilities are threatened. Further, the possible truth of determinism does not entail that surroundings are always hostile towards narrow abilities. So, determinism does not threaten narrow or (in some cases) wide abilities that constitute our freedom of will and action. A positive claim for free will is made based on possession of a bundle of intrinsic dispositions by creatures with minds and the ability to move their minds and bodies in goal-directed and other intelligent ways within surroundings that do not physically constrain actions (following Vihvelin 2013: 170 and 178).

Does Vihvelin make a compelling positive compatibilist case? The above is a sketch of arguments that take place over approximately two hundred and eighty pages[66] described by Vihvelin as ' … long, complicated, and indirect' (2013: 21). The key issue is whether such a model provides a convincing account of regulative control in addition to guidance control of conduct in a situation (world) where determinism happens to be true.[67] I am not going to attempt an analysis of Vihvelin's arguments, except to say that narrow abilities as bundles of dispositions do appear to be immune from determinism: 'No one thinks an unstruck match lacks the disposition to light simply because determinism is true' (Franklin Evan 2013). A threat to free will occurs when the wide ability to do otherwise is impacted by obstacles preventing exercise of our narrow abilities. Determinism is not considered to be an obstacle, even though its role appears to be the same *as* an obstacle i.e., preventing attempts to do otherwise from being successful. There are two issues here; the whole notion of narrow abilities as bundles of dispositions, but perhaps more important, (following closely Franklin Evan 2013), is the assumption that while a feature that prevents attempting to do otherwise from being successful is an obstacle, a feature that prevents even an attempt to do otherwise (determinism) is not considered to be an obstacle. I cannot discern from Vihvelin's text an explanation for this apparent problem.

[66] I have not mentioned Vihvelin's arguments and discussion relating to (i) what a successful defense of compatibilism entails, (ii) agent causation, and importantly, (iii) the claim that Frankfurt style cases are not genuine counter examples to the Principle of Alternate Possibilities.

[67] See Vihvelin and Fischer's discussion of this question in the 2008 edition of the *Canadian Journal of Philosophy* (Vihvelin 2008) and (Fischer 2008a) and for a more general critique see (Franklin 2013) or (Sartorio 2014).

2.2.2 Hierarchical Compatibilism

Part of Vihvelin's project within *Causes, Laws, and Free Will: Why Determinism Doesn't Matter* (2013) is a detailed critique of the claim that Frankfurt style cases offer genuine counter examples to the Principle of Alternate Possibilities, (see Chapter 2, page 43, for a brief description of Frankfurt's argument presented in *Alternate Possibilities and Moral Responsibility* (1969)). It is appropriate to continue by looking at hierarchical compatibilism via an aspect of Harry Frankfurt's compatibilism, his notion of freely willed actions originating from desires that *mesh* with important parts of a person's psychology, parts that are ordered or ranked in a hierarchical structure. This is a source model of control, whereby actions emanate from *the agent* rather than from any 'external' influence or control.[68] First, some essential concepts developed by Frankfurt:

First order desire: A desire to perform an action, i.e., a desire to go to the cinema.

Will: A first-order desire which is effective, i.e., that causes someone to do what they desire to do. A desire to go to the cinema is a person's will, in Frankfurt's sense, if that desire brings that person to actually go to the cinema.

Second-order desire: A desire to have a certain desire. A desire to go to the theatre (to see a culturally significant play) rather than the cinema (to see a trashy film) is an example of a second-order desire, (arguably, to fulfil a third-order desire, i.e., self-improvement).

Second-order volition: For someone to desire that a certain desire be their will, i.e., a desire that a certain desire brings a particular action. In terms of the above example, to have a second-order volition is to desire not just to have the desire to go to the theatre, but that the desire to go to the theatre rather than the cinema be effective in bringing about going to the theatre rather than the cinema.[69]

Frankfurt begins to describe his conception of free will in Part III of *Freedom of the Will and the Concept of the Person* using the concepts of first order desires, second order desires and particularly second-order volitions, (considered essential for personhood), claiming that 'it is only because a person has volitions of the second order that he is

[68] See also Appendix B Agency for further discussion of Agent causation.

[69] The 'cinema' example is from Jeff Speaks, Professor of Philosophy, University of Notre Dame, *Frankfurt's Compatibilist Theory of Free Will,* 19th March 2009. For a full explanation of the terms mentioned here see Harry Frankfurt, *Freedom of the Will and the Concept of the Person* (1971: 1-14).

capable both of enjoying and of lacking freedom of the will' (1971: 14), (another distinguishing mark of the human condition). Although the notion of being free as 'doing what one wants' is unclear in terms of the meaning of 'doing', 'wanting' and the relationship between these terms, Frankfurt does say something true about *acting* that is captured by this description. Freedom of will and freedom to act are separate in the sense that if an agent is deprived of the freedom to act (knowingly or unknowingly) the will to act is not thereby affected. Frankfurt describes freedom of the will as freedom '… to will what he wants to will, or to have the will he wants' (1971: 15). Willing freely happens when the will and a second-order volition are aligned.[70]

Earlier, a third-order desire of self-improvement was mentioned. Criticism of this model has been made based on ever-increasing levels of desires becoming too complex to be meaningful. Frankfurt recognises this point, saying common sense would prevail in terms of a self that would naturally limit the levels of active desire and volition. More convincing perhaps is the idea of a 'decisive' desire that 'rules them all' so no further questions and levels arise beyond the second level. Frankfurt makes some interesting comments about second-order volitions, essentially, they may not be deliberately formed, and the agent does not necessarily have to struggle to ensure they are satisfied. Frankfurt says:

> Examples such as the one concerning the unwilling addict may suggest that volitions of the second order, or of higher orders, must be formed deliberately and that a person characteristically struggles to ensure that they are satisfied. But the conformity of a person's will to his higher-order volitions may be far more thoughtless and spontaneous than this. Some people are naturally moved by kindness when they want to be kind, and by nastiness when they want to be nasty, without any explicit forethought and without any need for energetic self-control. Others are moved by nastiness when they want to be kind and by kindness when they intend to be nasty, equally without forethought and without active resistance to these violations of their higher-order desires. The enjoyment of freedom comes easily to some. Others must struggle to achieve it. (1971: 17)

[70] A clear and succinct description of this form of freedom is given by Daphne Brandenburg *Implicit Attitudes and the Social Capacity for Free Will* (2016: 1216).

Such comments clearly resonate with implicit bias, particularly '… without any explicit forethought and without any need for energetic self-control.'

On Frankfurt's account of freedom of the will and concept of a person, an agent acts freely when they do what they want to do and are free to want what *they* want to want. On this basis an agent having their *own* will experiences 'all the freedom it is possible to desire or to conceive' (1971: 17). How does this account relate to determinism, responsibility and the familiar claim that human freedom entails an absence of causal determination? An example of free action in a determined world is described based on the motion of a person's hand; when a person moves their hand, it is the outcome of a series of physical causes, but some event in this series, perhaps one of those that took place within the brain, was caused by the agent and not by *any* other event, (a source-compatibilist model). A free agent has, therefore, a 'God-like' quality, whereby they act as a prime mover unmoved (following Frankfurt 1971: 18). The idea of humans with 'God-like' qualities seems implausible and Frankfurt counters such a response by pointing out there is no difference between the *experience* of a man who miraculously initiates a series of causes when he moves his hand and a man who moves his hand without any such breach of the normal causal sequence. If there is no difference in *experience*, then there appears to be no reason to prefer being an unmoved-mover agent rather than a determined agent. Frankfurt essentially leaves the issue at this point and moves to a brief discussion of responsibility.

When discussing Compatibilism, (Chapter 2, page 42), a brief description was given of Frankfurt's argument against the Principle of Alternate Possibilities (PAP) based on his 1969 Paper *Alternate Possibilities and Moral Responsibility* (1969). Frankfurt's later Paper *Freedom of the Will and the Concept of the Person* (1971) also discusses PAP and for completeness I will briefly outline this discussion, containing the well-known 'willing addict' example of responsible action in the absence of alternatives. Frankfurt's ideas are relevant as they form part of what motivates John Martin Fischer's development of semicompatibilism.

Noting the historical link between theories of the freedom of the will and conditions of moral responsibility, and the recent approach whereby freedom of the will is considered *in terms of* what is entailed by the assumption that an agent is morally responsible, Frankfurt says:

> It is not true that a person is morally responsible for what he has done only if his will was free when he did it. He may be morally responsible for having done it even though his will was not free at all. (1971: 18)

In other words, this claim runs counter to the Principle of Alternate Possibilities (PAP) expressed in moral terms; that a person is morally responsible for their action if and only if that person could have done otherwise. Frankfurt's argument for this position is as follows. Suppose that an agent has freedom of action and freedom of the will, they are free to do what they want to do and free to want what they want to want. In other words, the agent could have chosen otherwise when deciding what they want to want. Frankfurt admits it is a 'vexed question just how "he could have done otherwise" is to be understood in contexts such as this one' (1971: 19), but claims this question is irrelevant when considering moral responsibility because, as claimed above, an agent may be morally responsible for an action even though their will was not free *in the sense that* alternatives that the agent opted against were not *actually* available. To illustrate an agent may be morally responsible for an action even though their will is not free in the above sense, Frankfurt uses the uncommon example of a willing addict, someone who would not change their situation and in the unlikely event of their addictive desire for drugs declining would take positive steps to mitigate the decline. The willing addict is not free, in that alternative drug-free lifestyle choices are not available because their addiction related desire and will to take drugs is present, active and irresistible whether or not they want this desire to constitute their will. The willing addict takes drugs freely and so responsibly in the sense that the action is supported by second-order desires and volition even though alternative choices are not actually available due to the irresistible force of addiction. Frankfurt describes this situation as an overdetermination of a first-order desire to take drugs: 'His will is outside his control, but, by his second-order desire that his desire for the drug should be effective, he has made this will his own. Given that it is therefore not only because of his addiction that his desire for the drug is effective, he may be morally responsible for taking the drug' (Frankfurt 1971: 20). Using a hierarchical mesh of first and second order desires and Frankfurt style scenarios (FSC), Frankfurt claims to show the possibility of freely willed (in the described sense) and morally responsible behaviour, even though the agent could not have done otherwise, hence refuting the traditional presentation of PAP. As noted previously (page 44, Footnote 56), Frankfurt offers an alternative version of PAP, taking into account issues raised earlier within *Alternate Possibilities and Moral Responsibility*; 'a person is not morally responsible for what he has done if he did it *only* because he could not have done otherwise' (1969: 838).[71] The importance of 'only' within this revised formulation is important in light of the 'willing addict' scenario, where responsibility is assigned on the basis that it is *not only* because of addiction that the willing addict takes drugs.

It is now approximately fifty years since the publication of *Alternate Possibilities and Moral Responsibility* and discussion of Frankfurt's arguments continues virtually unabated. I will note just two general responses that essentially claim that FSC's do not refute PAP because FSC's are not in fact cases where an agent is morally responsible for what they did and could not have acted otherwise. These responses are quite subtle and focus on the 'he could not have acted otherwise' part of the FSC. The first response is usually referred to as the flicker strategy because it is argued that within the FSC the agent does (must) experience a flicker of freedom to do otherwise. This flicker of freedom occurs at the instant before the controller reacts, or does not react, depending on whether the agent intends to act against or following the controller's wishes. The controller must receive by some means a sign that the agent intends or in some sense will try to act otherwise. There *must* be a moment when the possibility of acting otherwise triggers the controller's counter measure; the process of controller counter measure is at the heart of FSC. A simple FSC quickly shows the general problem: During a driving lesson I approach a particularly tight right-hand bend unaware that if I panic and begin to turn the steering wheel to the left by mistake the instructor will intervene. I panic, and in the instant before the instructor intervenes, I turn to the left by just the smallest degree, acting in a way that the instructor/controller is committed to prevent. In other words, it seems that I am not *totally* compelled (by the instructor), hence casting doubt

[71] Frankfurt's model within a world where determinism is true seems to me far less assured. For example, Frankfurt says; 'My conception of the freedom of the will appears to be neutral regarding the problem of determinism. It seems conceivable that it should be causally determined that a person is free to want what he wants to want. If this is conceivable, then it might be causally determined that a person enjoys a free will' (1971: 20). It is difficult to fully understand how this is possible, but it should not be overlooked that this account is essentially agent-source compatibilist in nature and further reflection on this basis may result in greater clarity.

on the legitimacy of the FSC and thereby its counter to the PAP. A reasonable response to this strategy is to say that while there may be such a flicker of freedom the agent does not (and cannot) actually act otherwise, (did my minute turn to the left or movement of my left hand constitute a turn or was it indicative of what I was about to do?), and so the flicker strategy is perhaps too weak to counter FSC.[72] As John Martin Fischer says, 'such a mere flicker of freedom would be too thin a reed to support the superstructure of moral responsibility' (2010a: 234).

The dilemma defence (DD) is the second argument designed to cast doubt on FSC as a legitimate counter to PAP. It is helpful perhaps, to keep sight of the bigger picture, to confirm the relevance of PAP and FSC to compatibilism, particularly semicompatibilism: It is good for compatibilists if FSC are resilient to DD[73], (or any counter), because if alternative possibilities are unnecessary for an agent

> to be morally responsible, then, arguments about whether causal determinism excludes the freedom to do otherwise, though perhaps interesting, become less important and even irrelevant to the question of whether causal determinism excludes *moral responsibility*. Even if causal determinism excludes the freedom to do other-wise, an agent might still be morally responsible because alternative possibilities are not necessary for moral responsibility (PAP is false). (added emphasis Widerker and Goetz 2013)

The DD argument will be outlined following a FSC created by John Martin Fischer (2010b: 319-320). Assume causal determinism does not obtain in a sequence of events that ends at t_2 with the agent Jones casting his vote. At an earlier time t_1 while reflecting on how to vote, Jones involuntarily shows a particular sign, such as a furrowed brow. When Jones furrows his brow in this way at t_1 he will reliably choose to vote Democrat at t_2. However, in a situation where determinism does not prevail, although voting Democrat *reliably* follows the earlier furrowed brow, given prevailing indeterminism, it is

[72] For an appropriately detailed discussion of these issues see, for example, Eleonore Stump, *Alternative Possibilities and Moral Responsibility: The Flicker of Freedom* (1999) and from Justin Capes, *The Flicker of Freedom: A Reply to Stump* (2014).

[73] For further discussion see David Widerker, Stewart Goetz, John Martin Fischer *Against the Dilemma Defence: The Defence Prevails* (2013), John Martin Fischer *The Frankfurt Cases: The Moral of the Stories* (2010b), John Martin Fischer *Frankfurt-Style Compatibilism* (2003) and David Palmer *Deterministic Frankfurt Cases* (2014).

possible that at t_2 Jones votes Republican. Black, the controller, makes no intervention, because the prior sign at t_1 leads Black to believe that Jones will vote Democrat as Black wants. It is argued that Jones could and did act otherwise even in the presence of Black. In other words, it is not true that Jones could not fail to vote Democrat, (he had an alternative possibility and took it by voting Republican), so this scenario does not counter PAP. Now, assume causal determinism does obtain in the sequence of events, independently of any intervention by Black. Within such a deterministic scenario it would surely be incorrect to claim that Jones acted responsibly on the basis that Black did not intervene; it would beg the question to assume that causal determinism is true *and* hold that it is uncontroversial to claim that Jones is morally responsible for his choice and action. Thus, again, the FSC counter to PAP is claimed to have failed as it does not present a convincing case of a morally responsible individual in the absence of alternative possibilities (following closely Fischer 2010b: 319-320).

The Dilemma Defence (DD) of PAP has generated much detailed and subtle debate. The very complexity of the DD detracts from its plausibility as a counter to FSCs. What exactly does '*reliably* follows' mean? The introduction of prior-signs and the notion of early decisions by Black at t_1 concerning possible intervention, are clearly very inventive, but do not I believe, lead to a *plausible* and robust counter to FSC's.[74] [75]

2.2.3 The Reason View

In Chapter 5 (page 144), referring to Susan Wolf, *Freedom within Reason* (1993), I briefly discuss the notion of 'who the agent is' in the context of Ecological Control, whereby implicit attitudes are regarded as part of an agent's character, part of 'who the agent is', and hence subject to moral evaluation. Wolf presents in *Freedom within Reason* a mesh theory between an agent's actions and their values. Here, an agent's actions are freely willed if they are in accordance with the True and the Good. As Michael McKenna and Justin Coates note, 'Because the conditions of Wolf's mesh theory require an anchor *external* to the agent's internal psychological states (the True and the Good), unlike

[74] For further discussion of the Dilemma Defence and other objections to Frankfurt Style Cases see Pablo Rychter *Does Free Will Require Alternative Possibilities?* (2017: 134-139).

[75] Clearly, inventiveness, for example, thought experiments involving brains in vats or asking the question what is it like to be a bat? *can* lead to great insights!

Frankfurt's, hers is not a real self-theory' (2015a). Following Wolf, explanation of the Reason View is helped by comparison with the Autonomy View and the Real Self View. The essential claim of the Reason View is that to act responsibly, (in significantly moral ways), an agent must act following reason rather than acting autonomously or in accordance with their Real Self.

Acting Autonomously and so responsibly is to act in a metaphysically distinctive way from the rest of nature. It is to act with the possibility to do otherwise, not determined, for example, by causal sequences beyond the agent's influence. The Reason View denies that responsibility rests on the possibility to do otherwise, to choose one action from several possible actions. Rather, it rests on the ability to act following reason, an ability that declares only one possible action and is not considered to metaphysically distinguish us from everything else in nature. We act responsibly by exercising our ability to reason, *thereby* recognising the True and the Good and then acting accordingly.

Followers of the Real Self View also claim that to act responsibly requires the ability to act in one way based on one principle, not considered to metaphysically distinguish us from everything else in nature; that principle is the ability to act following deep-seated values that are an expression of the agent's Real Self. Note the absence of *ultimacy* within these two views; it is the basis of control, reason or real self, which is fundamental. So, an agent acting following the Real Self View is responsible iff their actions reflect their substantial values. This is also true of those who support the Reasons View but with the added requirement that values are formed based on what is True and Good.

The Reasons View, by introducing value criteria into the definition of responsibility, (doing the right thing for the right reason), moves the focus away from problems inherent in the other two views, i.e., the metaphysical distinctness requirement of autonomy and the influence of external forces beyond the agent's control on the formation of the agent's real self that bring into question agent responsibility. (Problems associated with formation of the agent's real self are discussed later in the context of moral luck). Further, the Reasons View claims that the ability to do otherwise is unnecessary for moral praise and Wolf cites a supporting example of someone who sees a book for sale they know their friend would like; 'I had to buy it' they say, (doing the right thing, acting in accordance with the True and the Good, for the right reason), but such a comment does not lesson, and may actually increase, the recipient's positive feelings towards their friend (1993: 83). This account of responsibility raises many questions, for example, there is an issue of asymmetry within this model of freedom and responsibility. Praiseworthy conduct does not require the freedom to do otherwise (in that sense, it requires guidance control) but blameworthy behaviour does (regulative control) (Wolf 1993: 79). This apparently curious outcome of the Reasons View resonates with an idea first mentioned in Chapter One, that freedom is possible within a determined world by surrender of the will or aligning the will in some sense with the divine or the good. Here, the idea is expressed in terms of the agent acting in accordance with the True and the Good; if the agent is psychologically determined to act in accordance with the True and the Good, (the *defining* feature of freedom and responsibility), then being unable to act otherwise, (than in accordance with the True and the Good), does not threaten the sort of freedom that a morally responsible agent needs. If the agent is psychologically determined *not* to act in this freedom and responsibility characterising way, then being unable to act as reason requires would seem to rule out any blame for such actions. Wolf offers an explanation for this apparent asymmetry paradox (1993: 80), but there are more fundamental aspects of the Reasons View to consider; the Reasons View and determinism, and the nature of the True and the Good.

Using a story to develop and illustrate the relationship between the Reason View and determinism, Wolf concludes the *ability* of an agent that is necessary for responsibility is not *as such* incompatible with physical (or divine) determinism. Being physically impossible that an agent A perform an action X 'does not imply that A lacks the ability to do X, that A cannot do X, in any sense relevant to the assessment of X's responsibility' (Wolf 1993: 114). The foundation upon which this claim is built is a claimed distinction between physical determinism and psychological determinism; an agent is not determined at a psychological level, having the ability to choose one way rather than another, (for example, choice based on reason), with attendant responsibility. Essentially, the idea that physical determinism must interfere with psychological freedom is a mistake, or at least doubtful. As the Reasons View brings into play new and detailed ideas about objective moral values, free will, moral responsibility and God's foreknowledge an appropriate response would be quite extensive. Nevertheless, one important issue should be mentioned; on the Reasons View it is not completely clear how the agent is immune from artificial manipulation[76] in a way that undermines responsibility, (see Chapter 2, page 50). However, Wolf does address this point generally at the conclusion of the final chapter of *Freedom Within Reason* (1993: 142-147).

The Reasons View argues that responsible agents' actions are governed by reasons and knowledge of the world, where those reasons have as their locus agents' values, shaped by the True and the Good. True and good in the sense of clear, open-minded perception of the world that leads to true beliefs and good values rather than bad ones. The problem of objectivity of values is recognised, but Wolf makes the point that there are alternatives to established absolutist positions; what is claimed by the Reasons View is close to assumptions within ordinary moral conversations. The Reasons View takes an essentially objective position on values but, as suggested by the description Normative Pluralism, one that is clearly not absolutist, described by Wolf as 'objective enough' and 'partially objective' (1993: 126).

John Martin Fischer and Mark Ravizza comment 'Wolf's book is highly intelligent, original, and provocative' (1992: 389). An appropriately detailed criticism of the Reasons View is not possible here, but two issues suggest further reflection. First, the precedence of psychological explanation and psychological ability over physical determination when deciding if an agent could do otherwise in the relevant sense (with associated responsibility). Are psychological abilities of the sort described sufficient for freedom to do otherwise in the sense of guidance or regulative control and a convincing counter to incompatibilists like van Inwagen? In other words, the issue of ultimacy still seems to haunt this view, even though addressed by Wolf during the first three chapters of *Freedom Within Reason* (1993). Second, the asymmetry thesis is very counter intuitive. Recall, praiseworthy conduct does not require the freedom to do otherwise (guidance control) but blameworthy behaviour does (regulative control) (Wolf 1993: 79). As suggested by Michael McKenna and Justin Coates, the regulative control feature could

[76] Any agent who is determined to perform X is not different in any relevant respect from an agent manipulated into performing X. Therefore, when compatibilist's conditions are satisfied in cases of manipulation they could equally well be satisfied under conditions where determinism is true; such conditions then are not sufficient to guarantee *independence* from possible determinism with associated freedom and responsibility, (see Chapter 2, page 50). See also John Martin Fischer, *How Do Manipulation Arguments Work?* (2016).

perhaps be dropped as a condition of blameworthiness. Frankfurt-type examples are possible for cases of blameworthy action that seem to support this idea, i.e., neither praiseworthiness or blameworthiness require regulative control and so there is no asymmetry between the conditions of appropriate praise and blame. Preserving the asymmetry thesis requires regulative control for blameworthy conduct, with all of the issues that entails (following McKenna and Coates 2015a).

2.2.4 Reasons-Responsive Compatibilism

Regulative control allows the agent to regulate between different alternatives, by contrast, an agent acting with guidance control guides or brings about conduct, even if there are no other alternatives to the course taken. When used in the normal Fischer-Frankfurt sense, only guidance control is necessary for moral responsibility. Restating a key point, on this model, claims of compatibilist freedom do not include regulative control as illustrated by the Garden of Forking Paths. Other compatibilists may retain the classical compatibilist commitment to show that determined agents *can* act with regulative control (following McKenna and Coates 2015a: Section 5). John Martin Fischer's reasons-responsive compatibilism, semicompatibilism and implicit bias will be examined in Part III with the research aim of determining if behavioural expression of implicit bias *is* subject to the control conditions sufficient for an agent to assert guidance control i.e., is the issuing behaviour part of the agent's own moderately reasons-responsive actual-sequence mechanism? If behaviour that has implicit bias as its source *is* subject to guidance control, then the agent is responsible for such behaviour. If implicit bias related behaviour is 'beyond' guidance control, then the agent is not responsible. Reasons-responsive compatibilism will be described in greater depth in the next Chapter, Semicompatibilism.

2.2.5 Strawsonian Compatibilism

Strawson's *Freedom and Resentment* (1962) is mentioned in Footnote 59, page 45. In the following section Strawsonian compatibilism will be outlined with reference to this famous and influential Paper. At the heart of this compatibilist position is the notion of the moral community as the source of what it means to be an agent and whether agents are responsible for an act or behaviour. As McKenna and Coates describe, this is an 'anti-metaphysical reading' of the compatibilist position in the sense that

> it is the community that *constructs* a set of standards for freedom and responsibility that could be satisfied even in a determined world. Given that the conditions are constructed, they need not be constrained by prior metaphysical questions concerning the nature of the persons alleged to possess free will. (2015a)

In this radical approach, expressed simply, it is the community that decides what free will is, not something driven by metaphysical issues concerning the agent or the world generally.[77]

Strawson reminds us of the importance we attach to the perceived attitudes and intentions of other people when deliberating on how we feel about them. Our attitude towards others may be reactive, for example, resentment, or objective. An objective attitude is formed when, taking a step back from initial reactions, a better understanding of the situation is possible, for example, on reflection it is clear that harm was caused accidentally, or mental illness was instrumental in bringing about the harm. Strawson asks a question and develops an answer that summarises a significant part of his position:

> What effect would, or should, the acceptance of the truth of a general thesis of determinism have upon these reactive attitudes? (2008: 11)
>
> For it is not a consequence of any general thesis of determinism which might be true that nobody knows what he's doing or that everybody's behaviour is unintelligible in terms of conscious purposes or that everybody lives in a world of delusion or that nobody has a moral sense, i.e. is susceptible of self-reactive attitudes, etc. (2008: 19)

So, if determinism were shown to be an intrinsic part of our world should (rationally) the reactive attitudes be abandoned? No, because such reactive attitudes are too deeply ingrained within our humanity, and further, if it were possible to have a choice in this matter, then we 'could choose rationally only in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment; and the truth or falsity of a

[77] See Gary Watson's exceptionally clear and concise description of Strawson's argument in *Responsibility and the Limits of Evil: Variations on a Strawsonian Theme* (2013: 85).

general thesis of determinism would not bear on the rationality of *this* choice'(Strawson 2008: 14). That determinism is true is quite compatible with the existence of free will because free will is conferred by our reactive attitudes towards others.

There is a view (McKenna and Coates 2015b) that questions this model of compatibilism on the basis that it *includes* metaphysical considerations and assumptions. For example, it must be the case, it is claimed, that some metaphysical assumptions are made concerning the nature of agents that make them *proper* objects of our reactive attitudes and in an important sense make legitimate the position of those members of the moral community who display such attitudes. This view tends to push metaphysical issues of agency back towards centre stage. Strawson, I believe, would robustly resist this tendency, perhaps by emphasising and expanding on at least two points made within *Freedom and Resentment*. First, 'our practices do not merely exploit our natures, they *express* them' (2008: 27), and second, 'the existence of the general framework of attitudes itself is something we are given with the fact of human society. As a whole, it neither calls for, nor permits, an external "rational" justification' (2008: 25).

In this section I have described various Compatibilist positions: hierarchical compatibilism, the Reason View and Strawsonian compatibilism.

2.3 Hard Incompatibilism

Recall that essentially the libertarian position affirms that free will and determinism are incompatible and since human beings *do* have free will then determinism is false. However, the incompatibilist position also finds expression in denial of free will and acceptance of determinism. Such a position is usually referred to as hard determinism. This can be initially confusing; fundamentally, incompatibilism either denies determinism and asserts free will (libertarian), or denies free will and accepts determinism (hard determinism), (following Doyle 2017). In other words, determinism and free will cannot both be true at the same time. There is a further distinction to be made between *soft* and hard determinists. While both agree that our behaviour *is* determined, soft determinists are compatibilists, (in the sense that determinism does not undermine free will or responsibility worth having), whereas hard determinists are *in*compatibilists, (as described, determinism is true and free will does not exist in the sense required for genuine responsibility, accountability, blameworthiness, or desert) (following Kane 2012).

Bob Doyle, harshly perhaps, describes incompatibilism as a 'tortured and muddled concept […] a tar pit of confusion' (2017).[78] Doyle justifies his claim as follows:

> To be sure, *libertarians* have always denied the nonsense of compatibilism, and accepted the idea that free will is incompatible with determinism (*and we have free will*). Simple enough. But there is another view, that of *determinists* who agree that determinism is incompatible with free will. So, there are two kinds of incompatibilists, those who deny human freedom (usually called 'hard' determinists), and those who assert it (often called voluntarists, free willists, or metaphysical libertarians - to distinguish them from political libertarians). As a result, incompatibilism is a very confusing term in the free will debates. (added emphasis). (2017)

To arrive at *hard* incompatibilism consider the traditional and essential claims of hard determinism:

1. Free will is incompatible with determinism.
2. Free will does not exist (because)
3. Determinism is true.

Kane believes that today because of particular advances in theoretical physics commitment to proposition 3 has, to a large extent, fallen away and the incompatibilist position is now supported by proposition 1 and 2, described as the 'kernel of traditional hard determinism' (2015). Derk Pereboom calls this position hard incompatibilism; incompatibilist by proposition 1 and 'hard' by proposition 2. This section concludes with Derk Pereboom's description of hard incompatibilism. It is quite subtle, addresses several issues and is difficult to adequately summarise, hence reproduced in full:

> … I am agnostic about the truth of determinism. I contend, like Spinoza, that we would not be morally responsible if determinism were true, but also that we would lack moral responsibility if indeterminism were true and the causes of our actions were exclusively states or events. If the causes of our actions were

[78] A tar pit forms when subterranean bitumen leaks to the surface and creates a large and often deep area of natural viscous asphalt from which escape is usually impossible.

> exclusively states or events, indeterministic causal histories of actions would be as threatening to moral responsibility as deterministic histories are. At the same time, I think that if we were undetermined agent-causes – if we as substances had the power to cause decisions without being causally determined to cause them – we might well then have the sort of free will required for moral responsibility. However, although agent causation has not been ruled out as a coherent possibility, the claim that we are agent-causes is not credible given our best physical theories. Thus, we need to take seriously the prospect that we are not free in the sense required for moral responsibility. I call the resulting view *hard incompatibilism*. In addition, I argue that a conception of life without this sort of free will would not be devastating to morality or to our sense of meaning in life, and in certain respects it may even be beneficial. (Fischer et al. 2007: 85)[79]

Finally, in the next section, revisionism will be described.

2.4 Revisionism

> In a nutshell, revisionism is the view that what we ought to believe about free will and moral responsibility is different than what we tend to think about these things. (Fischer et al. 2007: 127)
>
> Revisionism about free will is the view that an adequate philosophical account of free will requires us to jettison some aspects of our common sense thinking about it. On this view, free will is like a host of other concepts, including scientific, moral, and conventional concepts which we have revised to more accurately reflect our understanding of the world. (Vargas 2008: 1)

Revisionism in the general historical sense is understood as rejection of traditionally held beliefs about a particular historical event or events (Collins online Dictionary 2020). The quotations above, from Manuel Vargas, loosely describe his revisionist position in the context of beliefs about free will.[80] Vargas' position is developed while looking at free will and responsibility from the point of view of common-sense intuitions and more rigorously by philosophical consideration; ' … our concepts of free will and moral

[79] See also Derk Pereboom *Free Will, Agency, and Meaning in Life* (2016), *Living without Free Will* (2003) and *The Significance of Free Will* (2000).

[80] See also Manuel Vargas *Revisionism about Free Will: A Statement and Defense* (2008). As noted in the Abstract of this Paper, Vargas 'summarizes and extends the moderate revisionist position put forth in *Four Views on Free Will* (Fischer et al. 2007) and responds to objections to it from Robert Kane, John Martin Fischer, Derk Pereboom, and Michael McKenna.'

responsibility should be revised in a way that renders them compatible with the natural physical order, even a deterministic one' (Fischer et al. 2007: 204). Vargas concludes that possession of free will, or acting from or with free will, occurs when an agent 'has the capacity to detect moral considerations *and* can govern him or herself (in the) appropriate way in light of those moral considerations' (Fischer et al. 2007: 160). In other words, to support responsibility an agent's freedom is sensitive to moral considerations and allows self-governance in light of such considerations. Free will exists in the satisfaction of these conditions. On his revisionist account Vargas says 'there is always more philosophical work to be done' (Fischer et al. 2007: 162) and continues, admitting some scepticism about being able to provide an account of a mechanism or faculty 'involved in the detection of moral considerations' and 'there being a single or general capacity for self-governance'. That said, as a revisionist account, a detailed case is made for addressing free will along 'leaner, revised lines' that offers the significant benefit of doing the work we require without the disconcerting difficulties entailed by the common-sense picture' (Fischer et al. 2007: 161).

2.5 Summary

I have briefly described the main contemporary positions and arguments concerning compatibility of freedom and determinism, giving further context and preparation for the next chapter and Part III. At the time of writing (September 2017), the web site 'Phil Papers' (<http://philpapers.org/>) holds over four thousand Papers within the category 'Theories of free will', (and there are twenty-two additional free will related categories), indicating the ever-growing volume and breadth of scholarship currently available. Although brief and descriptive in nature, the intention of this chapter has been to give some reference points, context and definitions that will be important when completing the main objective; to look at the meaning and impact of implicit bias on the semicompatibilist position.[81] In the next chapter I look specially at semicompatibilism.

[81] Excellent comparison of the main contemporary free will positions may be found in *Four Views on Free Will* (Fischer et al. 2007), where Fischer, Kane, Pereboom and Vargas describe their own positions and comment on their cowriters' contributions.

Chapter 3

Semicompatibilism

We live in the heroic age of free will philosophy.
Saul Smilansky[82]

3.0 Introduction

This chapter describes semicompatibilism in sufficient detail to take forward into Part III where the essential question, does implicit bias threaten the semicompatibilist position on free will and responsibility? will be addressed. Why choose semicompatibilism as the free will position to be examined in light of implicit bias? Semicompatibilism is probably the most popular compatibilist free will position, and its detailed arguments and claims invite an implicit bias centred critique that is missing from current literature. Another important reason is that John Martin Fischer's semicompatibilism is particularly *important* within contemporary discussion of free will. Introducing a special issue of *The Journal of Ethics* (Speak 2008) devoted to John Martin Fischer's *My Way* (2006a), Daniel Speak says 'John Martin Fischer's prolific work on the philosophical problems of free will and moral responsibility now spans three decades and its influence on these debates can properly be described as seminal […] his work has, after all, justifiably shaped the contemporary discussion of free will … ' (2008:123). In harmony with John Martin Fischer, Speak points out the assumption of personal moral responsibility is an essential feature of a person's self-image while living a normal moral life within society and should be resilient to how certain facts about the world turn out (2008:123). Given the importance, resilience, detailed exposition and, (perhaps most important), plausibility from a personal point of view, semicompatibilism is, I believe, the natural compatibilist choice for examination, in light of a possible threat from implicit bias. A challenge that in Part III semicompatibilism will be shown to resist.

[82] *Free Will and Moral Responsibility: The Trap, the Appreciation of Agency, and the Bubble* (Smilansky 2012: 212).

3.1 Semicompatibilism

John Martin Fischer's semicompatibilism has been described as 'the gold standard for cutting edge defenses of compatibilism' (McKenna and Coates 2015a); the semicompatibilist claims *moral responsibility* is compatible with the possible truth of causal determinism, even if causal determinism threatens regulative control (Fischer 2007: 71). John Martin Fischer develops two forms or types of control; guidance control that preserves moral responsibility (whether or not determinism is true) and regulative control:

> The semicompatibilist denies that the value of our free agency – or the basis of our moral responsibility – is the power to make a difference (to have regulative control). […] It may be that, just as there is a single line that connects the past to the present, there is only a single line into the future: a single metaphysically available path that extends into the future. In this case, what matters is how we proceed – how we walk down that path, (guidance control). […] For the semicompatibilist the basis of our moral responsibility is not selection in the Garden of Forking Paths, (regulative control), but self-expression in writing the narrative of our lives. (my additions are shown in parentheses). (Fischer 2007: 82)

There are many key works by John Martin Fischer, few better in terms of clarity than Chapter 2 of *Four Views on Free Will* (Fischer et al. 2007: 44-84), but to begin an outline of the basic argument and structure of John Martin Fischer's reasons-responsive compatibilism two précis by the original authors are useful concise sources; *Précis of Responsibility and Control: A Theory of Moral Responsibility* (Fischer and Ravizza 2000)[83] and *Précis of My Way: Essays on Moral Responsibility* (Fischer 2010a).[84]

First principles; moral responsibility is intrinsically linked with control. As described, there are two distinct types of control. Regulative control involves choice between alternative possibilities, guidance control does not. Frankfurt type cases present examples of how these two forms of control are separate and distinct; an agent can assert guidance control whether or not regulative control is possible. Importantly, although it is a plausible and widely held view that moral responsibility requires regulative control, Fischer and Ravizza argue that moral responsibility for actions, omissions, and

[83] Complete Work *Responsibility and Control: A Theory of Moral Responsibility* (Fischer and Ravizza 1998).

[84] Complete Work *My Way: Essays on Moral Responsibility* (Fischer 2006a).

consequences requires only guidance control. Moral responsibility does not require alternative possibilities, however, the traditional association of moral responsibility with control, albeit guidance control, is retained. The behaviour of an agent when acting with guidance control is via a reasons-responsive mechanism that is owned by the agent. It is claimed that 'an agent is morally responsible for an action, on our account, to the extent that this action issues from the agent's own, reasons-responsive mechanism' (Fischer and Ravizza 2000: 441). Immediately, the question arises, what exactly is a 'reasons-responsive mechanism' owned by the agent? Taking the second part first, 'owned by the agent' is defined as 'taking responsibility for behavior that issues from that kind of mechanism' (Fischer and Ravizza 2000: 441). More detail is provided by Fischer and Ravizza, who distinguish two kinds of context in which an agent might take responsibility for the kind of mechanism that leads to their behaviour. The formation of nonreflective and reflective attitudes are described that resonate with material within Chapter 4, particularly material relating to Dual Process/System theories of cognition. Nonreflective attitudes develop when a child is subject to reactive attitudes, education, parental guidance and perhaps punishment. The child naturally begins to have a sense of their own agency and to take responsibility for exercising that agency through mechanisms that deliver behaviour, including non-reflective habits. Greater reflection may lead an agent to question whether particular reactive attitudes are fair, even though society may consider them appropriate. The essential claim here is that an agent makes the mechanism that delivers their behaviour their own, (not as a result of manipulation for example), by taking responsibility for it (2000: 443). *Précis of My Way: Essays on Moral Responsibility* (2010a) improves the clarity of this concept, describing how

> one's mechanism becomes one's own in virtue of one's having certain beliefs about one's own agency and its effects in the world, that is, in virtue of seeing oneself in a certain way. … On my view, an individual becomes morally responsible in part at least by taking responsibility; he makes his mechanism his own by taking responsibility for acting from that kind of mechanism. In a sense, then, one acquires control by taking control. (2010a: 237)

As mentioned, guidance control has two components: the mechanism that issues in action or behaviour must be the agent's own mechanism[85] *and* the mechanism must be suitably reasons responsive. A particular version of reasons responsiveness is developed by John Martin Fischer called moderate reasons-responsiveness. The agent must be able to recognise reasons, some of which will be moral reasons, and react to reasons that are sound and sufficient for action. Further, it must be the actual sequence mechanism that is reasons-responsive for genuine moral responsibility.[86] Todd R. Long (2004: 151-172) summarises in a more formal way the notion of moderate reasons-responsiveness, reproduced below with minor changes in presentation:

An agent's responsibility relevant mechanism *K* is moderately reasons-responsive iff:

(1) *K* is regularly receptive to *reasons*, some of which are moral. This requires;
(a) That holding fixed the operation of a *K*-type mechanism, the agent would recognize reasons in such a way as to give rise to an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs.[87]
(b) That some of the reasons mentioned in 1(a) are moral reasons.

[85] It is especially important to note that Fischer and Ravizza's concept of moderate reasons-responsiveness that issues in action is always via a *different mechanism* from one that would operate if the agent's brain were manipulated in some way, for example by implants within the brain, where the agent is no longer morally responsible for the issuing actions because the mechanism is not their own. So, 'a brain-implanted mechanism that issues in an action would always constitute a different mechanism from an ordinary practical reasoning mechanism (owned by the agent) that issues in an action' (with my addition Long 2004: 151-172).

[86] John Martin Fischer clarifies; 'The actual sequence and the alternative scenario involve intuitively *different kinds of mechanisms*: in the actual sequence, there is the normal operation of the human capacity for practical reasoning, whereas in the alternative scenario there is significant and direct electronic stimulation of the brain by the neurosurgeon. Even though it is difficult to provide a general account of mechanism individuation, it is (in my view) intuitively clear that different kinds of mechanisms operate in the actual and alternative sequences of the Frankfurt-cases. Further, it seems to me that what grounds the moral responsibility of the agent in such cases are features of the *actual-sequence* (added emphasis) mechanism – properties of the path that actually leads to the behavior in question' (Fischer et al. 2007: 78).

[87] I believe this point addresses situations where 'weak reasons-responsiveness obtains by virtue of the agent's responsiveness to a bizarre reason, even though the agent is not responsive to a wide array of relevant reasons' (Fischer 2006a: 81).

(2) *K* is at least weakly *reactive* to reasons; this requires that the agent would react to at least one sufficient reason to do otherwise, (in some possible scenario), although it does not follow that the agent could have responded differently to the actual reasons.[88]

(3) *K* is the agent's own; being the agent's own means 'taking responsibility' for *K*. This requires that the agent;

(a) Sees herself as the source of her behavior (which follows from the operation of *K*).

(b) Believes that she is an apt candidate for the reactive attitudes that result from how she exercises her agency in certain contexts.[89]

(c) Views herself as an agent with respect to 3(a) and 3(b) based on her evidence for these beliefs.

It is possible to accept the conclusion of the Consequence Argument and accept that all causally deterministic sequences rule out regulative control, however, not all causally deterministic sequences pose problems for guidance control and therefore moral responsibility (following Fischer et al. 2007: 80). At the heart of this approach is the claim that the *actual*-sequence mechanism that issues in behaviour must be the agent's own and be moderately reasons-responsive; metaphysical issues about determinism are 'side stepped'. Crucially, moral responsibility is possible whether or not determinism is true.

In Part III, John Martin Fischer's reasons-responsive compatibilism will be studied with implicit bias to answer the question; is behaviour that originates or issues from implicit bias subject to *all* the control conditions that are necessary and sufficient for the agent to assert guidance control i.e., is the issuing behaviour part of the agent's own moderately reasons-responsive actual-sequence mechanism? Simply expressed, if behaviour that has implicit bias as its source *is* subject to guidance control, then the agent

[88] Item 2 is perhaps elusive to grasp; an example from John Martin Fischer helps: 'Consider my commendable act of working this afternoon for the United Way (see Footnote 124). Even though I would do so anyway, even if I had a publication deadline, I certainly would *not* work for the United Way if to do so I would have to sacrifice my job. Thus, the actual mechanism issuing in my action is weakly reasons-responsive' (2006a: 69). I believe this example and others from Chapter 3 of *My Way ~ Essays on Moral Responsibility* help to clarify.

[89] The 'reactive attitudes' reference relates to P. F. Strawson, *Freedom and Resentment* (1962), where Strawson calls the attitudes involved in moral responsibility the 'reactive attitudes'. See Section 2.2.5 Strawsonian Compatibilism, page 63.

is responsible for such behaviour. If implicit bias related behaviour is in some sense *'beyond'* guidance control, then the agent is not responsible. Guidance control bypasses various metaphysical issues concerning the compatibility of moral responsibility with causal determinism. However, does the notion of guidance control and semicompatibilism adequately and plausibly respond to the phenomenon of implicit bias?

3.2 Summary

John Martin Fischer's semicompatibilism is a reasons responsive compatibilism that rejects the key premise of the Classical Incompatibilist Argument (see page 38), that a freedom relevant condition for moral responsibility is the ability to do otherwise (regulative control). Sufficient freedom for moral responsibility is provided by guidance control. Guidance control requires the mechanism that issues in behaviour be the agent's own. It is an actual-sequence, mechanism-based, reasons-responsive form of responsibility enabling control. This actual sequence approach does not include any form of control involving genuine access to alternative possibilities during formation of the agent's character, the performance of actions by the agent (or choosing not to act) or bringing about consequences.

Incompatibilist manipulation cases challenge this (and some other) compatibilist positions, seeking to show manipulation of an agent's reasons responsive mechanism leads to their freedom and responsibility being plausibly undermined. It is argued that if this is possible, then determinism as manipulator also undermines an agent's freedom and responsibility. The semicompatibilist responds to the manipulation challenge in several ways, most effectively and fundamentally by arguing that a manipulated mechanism that leads to behaviour is not the agent's *own*, because they have not taken responsibility for it in the relevant way.[90] In Chapter 6 another possible control undermining problem will be discussed, the role of luck in the formation *of* an agent's mechanism.

In this chapter I have endeavoured to describe semicompatibilism in sufficient detail to take forward into Part III.

[90] See, for example, John Martin Fischer's Paper *Responsibility and Manipulation* (2004) for detailed discussion of semicompatibilism, responsibility and manipulation.

3.3 Summary of Part I

Chapter 1 aimed to answer the question, what is the (free will) problem? I believe, following Nagel (2003), 'the problem' is in essence the tension between, looking inwards at our self-image as agents in control of our lives, making responsible rational decisions and choices in a way that makes us feel like things are, in an important sense, 'up to us', and awareness of various challenges to such feelings of agency. Chapter 1 described how challenges to human freedom have taken many forms over time; the irresistible hand of fate and God's foreknowledge, science and the sense of an external world characterised by deterministic laws of nature. If 'the problem' is a clash of internal and external perspectives, then clearly one solution would be a decisive argument, proof or validation, that our sense of agency is in fact true, even considering so many sceptical challenges. Of course, an alternative solution would be a convincing argument or proof that human beings do *not* in fact enjoy such freedom, that our sense of agency and responsibility is an illusion. (The consequences of agency and responsibility being an illusion are controversial and the subject of much philosophical reflection). As well as these binary positions, there are clearly many others, such as compatibilism and its variations that seek to show freedom (in clearly defined forms) is compatible with the sceptical challenge of, for example, determinism. Some responses to 'the problem' are outlined in Chapter 2. The 'free will problem' connects with a large mesh of related issues, the most obvious and pervasive being moral agency, (human and nonhuman agency as a general concept), responsibility, praise and blameworthiness. There are other equally important connections too, for example, with political theory, psychology, criminal liability and punishment, philosophy of religion and metaphysical issues such as determinism.

From a broad consideration of historical responses to the free will problem attention focused on semicompatibilism, the free will position associated with John Martin Fischer. Chapter 3 explored semicompatibilism in more detail, with the aim of having sufficient clarity to take forward into Part III for critical examination considering the phenomenon of implicit bias. The reasons why semicompatibilism has been chosen as the free will position to be examined were described; essentially, because semicompatibilism is especially important, resilient and plausible. Semicompatibilism, described as the gold standard compatibilist position, captures our intuitions about agency and offers a rich and plausible model of our place in the world. John Martin Fischer believes his 'theory is cooler than some salient rivals' (Fischer 2012: 139) by sustaining our sense of moral responsibility and status as persons whether or not it happens to be the case that determinism, or indeterminism, is true. John Martin Fischer argues that semicompatibilism has the advantage of no hierarchical structures of consciousness with their associated difficulties and the ability to deal with more complex moral situations such as cases of weak will.

In summary, I believe Part I provides an outline of free will and semicompatibilism in sufficient detail for later critique as described. The way forward in Part II is to develop a similar understanding and presentation of implicit bias.

Having completed Part I and with a good sense of the aim of Part II, in anticipation of Part III, I will describe the possible outcomes of the exploration of implicit bias issuing behaviour and guidance control, confirming the way forward and destination of this Thesis. If implicit bias related behaviour is shown to be subject to guidance control, then implicit bias related actions are like other actions in the sense that we act freely and responsibly within the terms defined by semicompatibilism. On this outcome, findings are in harmony with the models of implicit bias developed in Part II, based on evidence and argument supporting agent responsibility for implicit bias related actions. If implicit bias related behaviour is shown not to be subject to guidance control, then such actions are not freely and responsibly conducted, in conflict with the developed models of implicit bias. Such conflict would show something is wrong, perhaps a deficiency within semicompatibilism that must be addressed. This is the essential activity of Part III. Careful examination of semicompatibilism and implicit bias will ultimately show that implicit bias related behaviour *is* subject to guidance control and agent responsibility, therefore in agreement with the developed models of implicit bias.

Part II

Implicit Bias

Chapter 4

The Origin and Meaning of Implicit Bias

> *Implicit bias is not a new way of calling someone a racist. In fact, you don't have to be a racist at all to be influenced by it. Implicit bias is a kind of distorting lens that's a product of both the architecture of our brain and the disparities in our society.*
> Jennifer Eberhardt[91]

4.0 Introduction

The opening sentences of Professor Gregg D. Caruso's Paper *Consciousness Free Will, and Moral Responsibility* (2016) from the *Routledge Handbook of Consciousness*, position the role of implicit bias within the free will debate as very much a live concern:

> In recent decades, with advances in the behavioural, cognitive, and neurosciences, the idea that patterns of human behaviour may ultimately be due to factors beyond our conscious control has increasingly gained traction and renewed interest in the age-old problem of free will. [...] Are agents morally responsible for actions and behaviours that are carried out automatically or without conscious control or guidance? (2016: 1)

The recent phenomenon of implicit bias in some important circumstances appears to threaten claims that human beings act freely and so responsibly. Under the influence of implicit bias an agent's actions in ethically relevant circumstances are influenced by factors widely claimed to operate below the radar of consciousness.[92] If an agent is

[91] *Biased* (Eberhardt 2019: 6).

[92] Possible meanings of 'below the radar of consciousness' will be examined later in this chapter. See also Greg Caruso *Précis of Neil Levy's Consciousness and Moral Responsibility* (2015).

unaware of influencing bias and prejudice it is argued that some actions are *thereby* not freely chosen because the agent is not *consciously* aware of all relevant factors affecting their decision-making; the agent lacks control in important responsibility reducing ways. It is commonly claimed that almost all of us to a greater or lesser extent are influenced by implicit bias manifest as 'automatic associations that influence decision-making often in negative ways, particularly judgment and evaluation of existing stereotypes or stigmatized groups' (Brownstein and Saul 2016a: 1).[93]

That inequalities exist within our society, perhaps all societies, is clear. The reasons are varied and often controversial; historical legacy, cultural, occupational, legal and explicit prejudice to name a few. This is a particularly poignant reality given that concepts of fairness, justice and the basic egalitarian principles of equality, equal rights and opportunities are held in high esteem by most of the population within liberal democracies. Such principles and beliefs are genuinely held and questioning their authenticity would probably cause great offence. However, as Michael Brownstein and Jennifer Saul point out, recent psychological research has shown that most people possess implicit biases that run counter to some of their explicitly held egalitarian principles (2016a: 1). For the majority, who believe they act following strongly held principles, it is shocking to hear that in practice their actions are, to a greater or lesser extent, influenced by ideas largely outside of conscious awareness and control, ideas that are explicitly rejected in conscious deliberation and conversation.[94]

[93] *Implicit Bias and Philosophy* in two volumes, edited by Michael Brownstein and Jennifer Saul (2016a and 2016b) provide perhaps the most substantial single source of implicit bias related material within philosophy.

[94] Speaking on NPR (a U.S. multimedia news organization) Mahzarin Banaji, co-author with Anthony Greenwald of *Blindspot: Hidden Biases of Good People* (Banaji and Greenwald 2013), describes the theory they worked on in the 1990's known as implicit bias. During the interview transmitted on the 17th October 2016 Mahzarin Banaji describes the moment she realised that our decisions are guided by forces we are not aware of: 'So just to go back a little bit to the beginning, in the late 1990's I did a very simple experiment with Tony Greenwald in which I was to quickly associate dark-skinned faces - faces of black Americans - with negative words. I had to use a computer key whenever I saw a black face or a negative word, like devil or bomb, war, things like that. And likewise, there was another key on the keyboard that I had to strike whenever I saw a white face or a good word, a word like love, peace or joy. I could do this very easily. But when the test then switched the pairing and I had to use the same computer key to identify a black face with good things and white faces and bad things, my fingers appeared to be frozen on the keyboard' (NPR 17th October 2016).

Considering Freud and others, the idea that unconscious factors influence behaviour is not new or expected to generate heated debate, however, the claimed pervasiveness, and more importantly, the disturbing nature and impact of implicit bias, is understandably an emotionally charged and hugely concerning issue.[95]

The vital question to be addressed in this chapter is what is the nature of implicit bias? I will begin with some foundations on which to build a response to this question. Beginning with description of the Implicit Association Test (IAT), I will then examine Dual Process and Dual System theory and cognitive models as an attractive explanatory paradigm for the influence of implicit bias in daily life.[96] Following these preliminaries, I will investigate implicit bias in terms of its nature and in Chapter 5 the importance of implicit bias within the free will and control debate will be considered.

4.1 The Implicit Association Test (IAT)

It is reasonable to ask, what is the evidence for the important and controversial claims described in the introduction to this chapter? Michael Brownstein describes as '… profoundly new … the ability to *measure* them (hidden prejudices) scientifically' (added emphasis 2016a: 3). Such methods are designed to avoid problems caused by asking

[95] Concern *and* practical response to implicit bias is manifest within Education, Criminal Justice, Health Care and Industry. For example, at the time of writing (May 2018), Starbucks US will close 8,000 coffee shops at 14.30 on the 29th May for an afternoon of anti-bias training for 175,000 employees in response to what is generally considered to be a racially motivated incident and the arrest of two customers in Philadelphia US.

See *The Telegraph* on line <https://www.telegraph.co.uk/news/2018/05/29/starbucks-close-8000-coffee-shops-us-racial-bias-training/> Telegraph Media Group Limited 2018. For US media comment on Starbuck's initiative see for example <https://www.vox.com/science-and-health/2018/4/19/17251752/philadelphia-starbucks-arrest-racial-bias-training>. Also Jennifer Eberhardt's discussion of this event in the context of implicit bias, *Biased* (Eberhardt 2019: 276).

The case of Jahi McMath and the death of George Floyd on the 25th of May 2020 in Minneapolis, takes consideration of implicit bias into quite a different and tragic realm. They are tragic examples of possible implicit bias within American Health Care, described by Michele Goodwin in her Paper *Revisiting Death: Implicit Bias and the Case of Jahi McMath* (2018), and possible implicit bias influenced behaviour of American Police Officers during an arrest for allegedly using a counterfeit bill.

[96] There is an obvious danger of circularity that should be kept in mind, in the sense that while the Dual System cognitive model appears a very appealing foundation upon which to build a description of implicit bias, any chosen model must of course be justified, as the form of model will have an impact on final conclusions about the nature of implicit bias that could be interpreted as being supportive of the Dual System model upon which such conclusions are based.

participants direct, often controversial, questions about their attitudes, for example, on ethnicity and gender issues. The goal of indirect methods is not to alert or inform the participants of what is being measured hence countering reflective responses and so (it is intended) accessing 'deeper' implicit attitudes. The IAT is designed to by-pass what participants believe they should say (self-presentation strategies), what is considered socially acceptable or conventional and what they believe they believe to reveal deeper attitudes that may be hidden from conscious awareness.[97] During the test the participant or subject is presented with a carefully developed series of images displayed on a computer screen that look typically as shown in Figure 4.1 A conventional keyboard is used as quickly as possible while trying to avoid error.

Black	White
Latonya	
Press key A to classify as Black or 5 to classify as White	

Figure 4.1 A typical IAT image from Sequence 1

A complete IAT schedule showing information about each sequence of images is shown below (Figure 4.2), reproduced from *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test,* by Anthony G. Greenwald, Debbie E. McGhee and Jordan L. K. Schwartz (Greenwald, McGhee and Schwartz 1998).

Sequence	1	2	3	4	5
Task Description	Initial Target Concept Discrimination	Associated Attribute Discrimination	Initial Combined Task	Reversed Target Concept Discrimination	Reversed Combined Task
Task Instructions	• BLACK WHITE •	• Pleasant Unpleasant •	• BLACK • Pleasant WHITE • Unpleasant •	BLACK • • WHITE	BLACK • • Pleasant • WHITE Unpleasant •

[97] For a detailed description of this point and IAT's generally see *A Practical Guild to IAT's and Related Tasks* (Teige-Mocigemba, Klauer and W. Shermen 2010: 117).

Sample Stimuli	MEREDITH o o LATONYA o SHAVONN HEATHER o o TASHIKA KATIE o BETSY o o EBONY	o Lucky o Honor Poison o Grief o o Gift Disaster o o Happy Hatred o	o JASMINE o Pleasure PEGGY o Evil o COLLEEN o o Miracle o TEMEKA Bomb o	o CORTNEY o STEPH SHEREEN o o SUE TIA o SHARISE o o MEGAN NICHELLE o	o Peace LATISHA o Filth o o LAUREN o Rainbow SHANISE o Accident o o NANCY

Figure 4.2 A Complete IAT Schedule

The A-key is pressed using the left hand to select the upper left category or the 5-key using the right hand to select the upper right category. For example, in Sequence 1 choices are made by this method between Black (upper LHS) or White (upper RHS) known as the target concepts. Choices are made in response to a sequence of names, for example, Shavonn, Heather, Tashika or Katie. The correct categorisation is shown by an open dot to the right or left of the name in the schedule, (see Figure 4.2). There are five sequences, a solid dot shows the category appearing at upper LHS of the displayed image and likewise a solid dot to the right of the category appearing at upper RHS. Crucially, the time taken for the subject to press the chosen key after each of the images is presented is recorded. In Sequence 3 and 5 the target-concepts Black and White are presented together with the attributes Pleasant and Unpleasant, following the solid dot LHS – RHS convention. A typical image from sequence 3 is shown in Figure 4.3 Note the 'or' between Black and Pleasant on the LHS and between White and Unpleasant on the RHS. The subject must decide if the sample stimuli 'Latonya' a typically black African American name (from circa 1998) is associated with Black or Pleasant, or, White or Unpleasant. Clearly, the correct response is LHS A-click.

Black **or** Pleasant	White **or** Unpleasant
Latonya	
Press key A to classify as Black or 5 to classify as White	

Figure 4.3 A typical IAT image from Sequence 3

The complete test, Sequence 1 through to 5, is conducted as follows:

Sequence 1: Introduction of target-concept discrimination.[98] This first discrimination distinguishes first names that are (United States circa 1998) recognizable as Black or African American from names recognizable as White or European American.

Sequence 2: Introduction of the attribute dimension, also in the form of a two-category discrimination. The attribute discrimination is an evaluation, categorizing words as having a generally pleasant or unpleasant meaning.

Sequence 3: After introduction of the target discrimination and the attribute dimension the two are superimposed in the third step in which stimuli for target and attribute discriminations appear on alternate trials. Figure 4.2 quickly clarifies.

Sequence 4: In the fourth step the respondent learns a reversal of response assignments for the target discrimination.

Sequence 5: The fifth and final step combines the attribute discrimination (not changed in response assignments) with this reversed target discrimination. Again, easy clarification may be found with reference to Figure 4.2

The objective of the IAT is to 'assesses the association between a target-concept discrimination and an attribute dimension' (Greenwald, McGhee and Schwartz 1998: 1465). At the heart of the test is the principle that if the target categories are differentially associated with the attribute dimension, the subject should find one of the combined tasks (the third or fifth step) to be considerably easier and hence completed faster than the other.[99] Further, and perhaps most important, the measure of this difficulty difference provides the measure of implicit attitudinal difference between the target categories (Greenwald et al. 1998: 1466). For example, for White subjects raised in a culture where anti-Black discrimination persists, the subject is expected to find choosing 'Black or Pleasant' harder (slower) than choosing 'White or Pleasant', see Figure 4.3. Attitudes are claimed to be revealed because associations running without active thought

[98] Clearly, the term 'discrimination' is used here in the sense of recognition and understanding of the difference between one thing and another, specifically, the ability to distinguish between different stimuli.

[99] Gaertner and McLaughlin's *Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics* (1983) is regarded by some as the first article to claim demonstration of implicit stereotyping. This Paper argues that unrelated to their degree of explicit prejudice, subjects 'responded reliably faster when positive attitudes were paired with Whites than with Blacks or with Negroes' (1983: 23).

(automatically) relatively help performance in one of the IAT's two combined tasks. Participants in the IAT experience a higher, (conscious, controlled, explicit, reflective, analytic, rational), level of mental operation when trying to *overcome* the effects of automatic associations (Greenwald et al. 1998: 1466). The level of mental operation is reflected proportionally in the response time to each image, example results are shown in Figure 4.4 from Greenwald's Paper (1998: 1474).

This Figure is available from *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test* (Greenwald et al. 1998: 1474)

Figure 4.4 Response Times for each IAT Sequence[100]

Greenwald and his team were confident that their experimental data

> clearly revealed patterns consistent with the expectation that White subjects would display an implicit attitude difference between the Black and White racial categories. More specifically, the data indicated an implicit attitudinal preference for White over Black, manifest as faster responding for the White + Pleasant combination (white bars above) than for the Black + Pleasant combination (black bars). (1998: 1474)

[100] Information is presented in a 'box plot' format. This format usually depicts groups of numerical data through their quartiles. The lines extending vertically from the top of the boxes (whiskers) show the maximum values. There are numerous web sites that describe box plot theory, for example *Simple Psychology* <https://www.simplypsychology.org/boxplots.html> or in greater detail, *Towards Data Science* <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>. (The sequence in Fig 4.4 does not follow that of Fig 4.2 but the argument is not affected by this discrepancy).

The IAT is an indirect measure intended 'to reveal implicit attitudes which are hypothesised to underpin discriminatory behaviours' (Holroyd and others 2017: 2). As expected, there is *much* critical comment concerning the IAT in terms of its method and principle.[101] I describe the IAT because it is the leading, most well-known, implicit attitude evaluation procedure and naturally introduces exploration of models of cognition where differential speeds of response suggest distinct modes of information processing. Models that offer explanatory paradigms for the presence of attitudes developed in response to experience, no longer available to conscious memory, that continue to influence behaviour (Following Brownstein and Saul 2016a: 8).

4.2 Duality of Mind: Dual Process and Dual System Models of Cognition

There is strong, many would say irrefutable, evidence that unconscious bias significantly influences important decisions for many people (J. A. Bargh 1999). During earlier discussion of implicit bias, the expression 'operating under the radar of consciousness' was used, but loose description of this kind must be clarified in terms of the nature of human cognitive architecture that engenders or allows phenomena such as implicit bias to be present and influential. Mandelbaum expresses a similar point; 'The study of implicit bias is deeply intertwined with questions of how learning interacts with cognitive structure' (2016: 1). Dual Process and Dual System models offer such a rich descriptive paradigm of human cognition. While such theories of cognition look attractive, there must be careful investigation into the legitimacy of associating implicit bias and this type of model.

There is a vast body of work available on Dual Process and Dual System models, therefore selecting Papers within a brief overview inevitably omits some important material. However, to engage appropriately with implicit bias it is necessary to give a limited introduction to Dual Process and Dual System models. The order and titles of the Papers listed below indicate the shape and content of what is to follow:

- *The Duality of Mind: An Historical Perspective* (Evans and Frankish 2012).
- *Implicit Learning and Tacit Knowledge* (Reber 1989).

[101] There is a considerable body of critical work concerning the IAT. For examples, see *Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT* (Blanton et al. 2009) and *A Closer Look at the Discrimination Outcomes in the IAT Literature* (Carlsson and Agerstrom 2015).

- *Evolutionary Versus Instrumental Goals: How Evolutionary Psychology Misconceives Human Rationality* (Stanovich and West 2003).
- *Heuristics and Biases; The Psychology of Intuitive Judgement* (Gilovich, Griffin, and Kahneman 2002).
- *Judgment under Uncertainty: Heuristics and Biases* (Tversky and Kahneman 1974).
- *Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition* (Evans 2008).
- *Dual Process Theories* (Gawronski and Creighton 2013).
- *Stereotypes and Prejudice: Their Automatic and Controlled Components* (Devine 1989).
- *A Perspective on Judgment and Choice* (Kahneman 2003).
- *The Nature of Intuitive Thought* (Jarvilehto 2015).

Dual Process and Dual System models and their related theories are based on the essential claim that the mind processes information in a way that is either[102] (i) fast, automatic, nonconscious, associative, responsive to experience and slowly acquired by social conditioning (Type One process), or (ii) slow, controlled, conscious, rule governed and capable of learning in response to explicit tuition (Type Two process), (following Brownstein and Saul 2016a: 10). Description of this paradigm of cognition begins with reference to *In Two Minds: Dual Processes and Beyond* (Evans and Frankish 2012). The first chapter presents a strikingly clear and detailed historical perspective of duality of mind. First, it is important to clarify the terms Dual Process and Dual System. From the 1970's Dual Process theories have been developed that focus on various aspects of human psychology, for example, deductive reasoning, decision making and social judgment. The common feature of these theories is the presence of two distinct processing mechanisms *for each task* that may yield different, and sometimes conflicting, results. As Evans and Frankish note, 'typically, one of the processes is characterized as fast, effortless, automatic, nonconscious, inflexible, heavily contextualized and undemanding of working memory, and the other as slow, effortful, controlled, conscious, flexible, decontextualized, and demanding of working memory' (2012: 2). More recently, much

[102] There are many different models, many considerably more sophisticated than this binary structure. See for example, *Dual Processing Accounts of Reasoning, Judgment, and Social Cognition* (Evans 2008), *Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems* (E. Smith and De Coster 2000) and particularly *Automatically Minded* (Fridland 2017). Fridland argues in Section 3 that 'at least some automatic processes are likely cognitively penetrable', see also Fig 4.5 below.

effort has been made to unify such Dual Process theories into a theory of mental architecture whereby human central cognition is composed of two multi-purpose reasoning *systems*, usually referred to as System One and System Two (Evans and Frankish 2012: 2), (Gawronski and Creighton 2013: 296). The former having fast-process characteristics, variously described as effortless, automatic and nonconscious, the latter a slow-process system, described typically as requiring greater mental resources, voluntary and conscious. Such unified architecture, forming System One and System Two, is unsurprisingly often referred to as a Dual System theory, in contrast to more localized Dual Process theory.

A history of the idea of the mind as divided in function or nature begins as early as Plato, where soul (or mind) is divided into three parts: reason, spirit, and appetite, each having an ability to influence action according to their own goals and powers.[103] While it is tempting to detour into description of how ideas of duality of mind have developed since Plato, (the well-known dual mind theorist Sigmund Freud was mentioned briefly in Chapter 1), this outline will be restricted to later developments of Dual Process and System approaches as presented within some key Papers. To confirm, processes characterised as either fast and automatic (Type 1) or slow and deliberative (Type 2) may be unified within Dual System theories that attribute the origin of these processes to two distinct cognitive systems. Frankish shows accumulated characteristics of each System gathered from various writers and reproduced below in Figure 4.5 (2012: 21), crediting A. S. Reber, particularly his 1989 Paper *Implicit Learning and Tacit Knowledge* (1989), as a major and early contributor to this characterisation.

[103] See particularly Book IV of the *Republic*, where Socrates and his interlocutors, Glaucon and Adeimantus, try to answer the question, does the soul consist of one part or several parts?

System One	**System Two**
Evolutionary old	Evolutionary recent
Unconscious / preconscious	Conscious
Shared with animals	Uniquely human
Implicit knowledge	Explicit knowledge
Automatic	Controlled
Fast	Slow
Parallel	Sequential
High Capacity	Low capacity
Intuitive	Reflective
Contextualised	Abstract
Pragmatic	Logical
Associative	Rule based
Independent of general intelligence	Linked to general intelligence

Figure 4.5 Accumulated characteristics of System One and System Two

Reber's Paper extends and develops conclusions of his earlier work[104] arguing that

> implicit learning is characterized by two critical features: (a) It is an *unconscious* process, and (b) It yields abstract knowledge. Implicit knowledge results from the induction of an abstract representation of the structure that the stimulus environment displays, and this knowledge is acquired in the absence of conscious, reflective strategies to learn. (1989: 219)

In the 1989 Paper, Reber looks at empirical data gathered since earlier publications of the mid 1960's and presents an 'overview of this new evidence and attempts to extend the general concepts to provide some insight into a variety of related processes such as arriving at intuitive judgments … '(1989: 219). This Paper reemphasises the claim that

> … a considerable portion of memorial content is unconscious, and, even more important, a goodly amount of knowledge acquisition takes place in the absence to the intent to learn. (1989: 230)

[104] See for example Reber's 1965 unpublished M.A. Thesis *Implicit Learning of Artificial Grammars,* (while at Brown University) and the 1967 article *Implicit learning of artificial grammars* from the Journal of Verbal Learning and Verbal Behavior, Volume 77, pages 317-327. (Both of these Works are cited within *Implicit Learning and Tacit Knowledge*).

Reber continues, arguing that unconscious cognitive processes should receive more attention, that focus on the conscious over the unconscious is to incorrectly prioritise what is most important, noting in evolutionary terms that consciousness is 'a late arrival on the mental scene' (1989: 230). The essential claim is that unconscious mental processes form the crucial foundation upon which conscious processes emerge, however, with the emergence of consciousness the capacities of the unconscious are in no sense diminished. This raises the obvious but unanswered question; how do these modes of cognition interact, for example, does one mode have ultimate control over the other?

Interestingly, Reber defines two subcategories of the *unconscious*: the primitive and the sophisticated. Operations of the primitive unconscious concern the most fundamental necessities for survival of living nonvegetative organisms. Operations within the realm of the sophisticated unconscious depend upon knowledge acquired without conscious awareness by primitive processes *and* act causally 'to control perception, affective choice and decision making independently of consciousness' (1989: 232). This is a large claim and is followed by another; that sophisticated systems are *available* to consciousness. This is a confusing claim, as earlier both 'primitive' and 'sophisticated' systems were described as subcategories of the unconscious. I believe Reber's text to be unclear on this point, but suggest the point being made is that while such sophisticated knowledge is acquired and forms the basis of some actions unconsciously, the subject is in an undefined way aware of such knowledge. This is markedly different to implicit bias (as widely presented[105]) where the subject strongly and truthfully denies any knowledge (awareness) of, for example, negative beliefs about particular ethnicities, religious communities and so on. The whole confusing and controversial issue of implicit bias as possible knowledge, belief, attitude, or intuition will be considered later in this Chapter. Reber concludes by summarising along lines already described, remarking that the key problem, one that is also of immense importance in the context of implicit bias, is to

[105] Bertram Gawronski comments, when endorsing *An Introduction to Implicit Bias* (Beeghly and Madva 2020) 'The science of implicit bias is rather complex – much more complex than suggested by the dominant polarized views in the public discourse.'

> specify, as clearly as possible, the boundary conditions on the process of implicit learning, that is, to outline the circumstances under which it emerges and those under which it is suppressed or overwhelmed. (1989: 233)

The terms 'implicit learning' and 'implicit knowledge' used by Reber and elsewhere must be considered carefully when used during discussion of implicit bias. It is not the purpose of Reber's Paper to be involved in a philosophical discussion of the nature of knowledge and implicit bias cannot be regarded *as* knowledge when understood in terms of the classic conditions of justified true belief. In fact, implicit biases may be considered 'always epistemically *bad* if we adopt a fog metaphor' (Beeghly 2020: 77) and their badness is multidimensional. Importantly, 'Biases are widely thought to articulate false or misleading claims about groups, which - once internalized - taint perceptual and cognitive judgments about individuals' (Beeghly 2020: 81).

Returning to discussion of the development of Dual Process and System models, Keith Stanovich and Rich West's Paper *Evolutionary Versus Instrumental Goals: How Evolutionary Psychology Misconceives Human Rationality* (2003) [106] is a challenging and important work. It is important in the context of this Thesis because it presents a clear argument supporting a Dual System paradigm of cognition, significantly, a paradigm that appears to offer an attractive explanatory model of implicit bias.

I will now describe the Development of a *unified* Dual System model, beginning with Stanovich and West and their argument

> that Dual Process[107] models of cognitive functioning provide a way of reconciling the positions of the evolutionary psychologists and researchers in the heuristics and biases tradition. (2003: 2)

To see how this argument develops, a brief explanation of the heuristics and biases tradition and the evolutionary psychologist position is needed. For insight into the

[106] Page numbers within citations for this work refer to the Paper available on web page <https://semioticon.com/virtuals/imitation/kstanovich_paper.pdf> rather than Chapter 7, *Evolution and the Psychology of Thinking: The Debate*, edited by D. E. Over.

[107] As described, it is important to keep in mind the distinction between Dual Process and Dual System. Although Stanovich and West use the term *Process* here, it will be seen that reconciliation is achieved by a Dual *System* approach.

heuristics and biases tradition there is probably no better source than *Heuristics and Biases: The Psychology of Intuitive Judgement* edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman (2002) and the earlier *Judgment under Uncertainty: Heuristics and Biases* by Amos Tversky and Daniel Kahneman (1974).[108] The Introduction to *Heuristics and Biases: The Psychology of Intuitive Judgement* describes the 'central idea of the "heuristics and biases" program' as follows. Judgements made 'under uncertainty often rest on a limited number of simplifying heuristics rather than extensive algorithmic processing' (Gilovich et al. 2002: 1). The influential work *Judgment under Uncertainty: Heuristics and Biases* argues that the processes of intuitive judgement are not just simpler than those of rational models but are categorically different. Three heuristics are proposed that under uncertainty would be used to make intuitive judgements. Although categorically different such heuristics 'piggyback' (using Gilovich and Griffin's expression) basic evolved mental processes. Each heuristic has an associated set of biases that function as markers, showing that a certain heuristic has been used.

An example from *Judgment under Uncertainty: Heuristics and Biases* illustrates the point with reference to a particular heuristic and bias. Questions raised in everyday life are often of the form, what is the probability of an object A belonging to a class B? The heuristics and biases paradigm would describe responses to such questions as 'typically relying on the 'representativeness heuristic' where probabilities are evaluated by the degree to which A is representative of B', the degree to which A *resembles* B. If A closely resembles members of class B, then it is thought to be highly and decisively probable that it is a member of class B. If someone is described by terms such as tidy, shy, helpful with a need for order and structure, how would the probability of being engaged in a particular occupation be assessed from a list of possibilities that include pilot, farmer, sales representative or librarian? From the Heuristics and Biases account it is highly likely that the representativeness heuristic would be employed. The highest probability would in most cases be assigned to the occupation of librarian, the assessment made by the degree to which the subject is representative of, or like, the librarian stereotype. However, judgements made about probability using similarity or representativeness

[108] See also *Heuristics and Biases: Beyond Tversky and Kahneman's (1974) Judgment under Uncertainty* (Fiedler and Sydow 2015) for a critique of Tversky and Kahneman's Paper.

often led to serious errors because similarity, or representativeness, is not influenced by factors that should be considered when assessing probability. Tversky and Kahneman describe one such factor as 'insensitivity to prior probability of outcomes' (1974: 1124). This is easily understood with reference to the current example; there are more farmers than librarians within the population and this should be considered when estimating the probability that the subject is a librarian, rather than a farmer. Prior probability of outcome does not affect the similarity of the subject to the stereotypes of librarian, or farmer.

Other heuristics and associated biases have been proposed and have generated much critical literature.[109] Due in large part to the brevity of the above description the heuristics and biases model may not appear particularly revolutionary, perhaps just an expression of common sense. However, as Gilovich and Griffin point out, the heuristics and biases paradigm questioned the adequacy of ideal models of judgement that describe rational people making choices based on sound assessment of the probability of alternative outcomes, the utility derived from each and combining these assessments in the decision-making process. The heuristic and biases model challenges the idea of the ideal rational agent, raising doubts about the competence of agents to make complex entirely rational calculations. The model of the rational agent was pervasive, particularly within the discipline of economics and casts a large shadow over any discussion of the modern history of research on everyday judgements [110] (Gilovich et al. 2002: 1), (Kahneman 2003: 702), (Kahneman 2003: 705).

A brief explanation of the heuristics and biases tradition and an outline of the evolutionary psychologist's position is necessary to show how Dual System models of cognitive function offer reconciliation of these two positions, forming the dominant

[109] See also Sarah-Jane Leslie's fascinating Paper, *The Original Sin of Cognition: Fear, Prejudice, and Generalization*. While not a heuristic, our primitive tendency to quickly generalise strikingly negative information across members of highly essentialized groups, leading to prejudice and serious errors of judgement, does resemble the operation of heuristics (2017: 421).

[110] Considering the central theme of this Thesis, there is a particularly interesting comment in the Introduction to *Heuristics and Biases: The Psychology of Intuitive Judgement*. Discussing how various factors came together to boost interest in Heuristics and Biases, Gilovich and Griffin comment, '… the greatest fascination for social psychologists has always been the combination of stereotyping, prejudice and discrimination, topics to which the heuristics and biases agenda was seen as highly relevant' (Gilovich et al. 2002: 7).

model of cognitive architecture. So, turning now to the evolutionary psychologist's position, this can be framed as a criticism of the heuristics and biases model. Gilovich and Griffin (2002: 9) call this criticism the 'people are not that dumb' critique, based on the claim that the heuristics and biases model is too pessimistic in its account of an agent's competence to make sound judgments. The evolutionary psychologist is sympathetic to this criticism in the sense that competence to make sound judgments is believed to be absolutely necessary for survival, hence evolution acts as an effective driver towards the ability to *make* sound, error free, decisions. The extent and diversity of human achievement shows quite obviously our ability to make good judgements. However, as Gilovich and Griffin point out, although initially appearing to be a sound claim, this argument on further reflection does not deliver a decisive blow to the heuristic and bias model. While 'evolutionary pressures lead to adaptions that are as good or better than a local rival they do not lead to adaptions that are globally optimal' (2002: 9). In other words, evolutionary pressures are not able to shape mental ability to the extent that error or bias free judgements are guaranteed.

As Gilovich and Griffin note, the heuristic and bias model has weathered several critiques and remains vigorous in part due to its relationship with psychology, particularly the great interest in 'automatic' mental processes. Dual Process and Dual System model development is linked *with* developments in the field of heuristics and biases, suggesting a symbiotic relationship between the disciplines. It is suggested this close relationship is more than a mutually beneficial development; the heuristic and bias model may usefully be expressed and discussed in the language *of* System One and Two architecture and Gilovich and Griffin offer good explanations of heuristics in terms of the System One and Two model (2002: 17).

Having made some brief explanatory notes on claims relating to heuristics and evolutionary psychology I will return to *Evolutionary Versus Instrumental Goals: How Evolutionary Psychology Misconceives Human Rationality* (Stanovich and West 2003) and the idea that Dual System models of cognitive function offer reconciliation of the evolutionary psychologist and heuristics traditions, (note Footnote 107). The heuristics tradition understands the results of many empirical demonstrations as evidence of human cognition characterised by systematic irrationalities. However, various evolutionary psychologists have interpreted these results quite differently, as indicating 'an *optimal* information processing adaptation on the part of the subjects' (Stanovich and West 2003: 1). How is this opposing interpretation to be understood? Stanovich and West bring to our attention the hypothesis that the driving force behind the development of primate intelligence was the need to master the social world.[111] The essential and important claim is that social interaction engenders a *particular form* of intelligence that acts as a foundation upon which 'future evolutionary and cultural developments in modes of thought are overlaid' (Stanovich and West 2003: 4). The social orientation toward problem solving is always available, it has not diminished over time and acts as a default resource. This leads to a key point in the developing argument; it is this long-standing foundation or substrate of social intelligence that responds when judgements are needed, providing the modal responses associated with the heuristics and biases model. Further, Stanovich and West argue, supported by substantial empirical data, that common (modal) responses based on judgement employing heuristics are often different to the judgements made by more cognitively able subjects. Less able subjects base their judgement on social cues, linguistic cues and background knowledge rather than the abstract reasoning of their peers. Importantly, such empirical data suggested to Stanovich and West that evolutionary rationality, based on social intelligence, is divided or separated from rationality employed by subjects that have higher cognitive ability, characterised by the ability to decontextualize and depersonalize problems, seeking out underlying principles and dealing with problems without the need for social content or conversational relevance.

Reconciliation of evolutionary and heuristic paradigms is brought about under a *unified* Dual System Model: System One; highly contextualized, personalized and socialized. The notion that evolutionary social intelligence will provide responses based upon heuristics that incorporate important relevant conversational cues and assumptions based on experience, and System Two; more controlled processes serve to decontextualize and depersonalize problems, displaying what has been described as analytic intelligence (following Stanovich and West 2003: 9). The obvious and important

[111] Reference is made to *The Social Function of Intellect*, an article by Nicholas Humphrey first published by the Cambridge University Press in 1976 and part of the collection *Growing Points in Ethology* edited by P.P.G. Bateson and R.A. Hinde.

question concerning the exact nature of the interaction between the two systems is still (temporally) unanswered. This is clearly an important issue in the context of implicit bias, as the *possibility* of overriding the influence of implicit bias, with associated implications for responsibility, appears to rely on System Two processes having some form of executive control of over-learned social norms characterised as being within the System One domain. Responsibility seems to be a function of the *extent* that System Two processes ultimately control behaviour. Such matters are clearly important and will be considered in detail shortly.

Before engaging in detail with control issues, I will give further confirmation of the relevance of Dual Process and System ideas to implicit bias and a summary of the way forward. This section on Duality of Mind began with reference to *In Two Minds: Dual Processes and Beyond* by Jonathan Evans and Keith Frankish and it is with reference to another work by Jonathan Evans, *Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition* (2008)*,* that the vital link between models of human cognition and implicit bias may be clearly seen. The importance of Dual Process and System theories of *social* cognition and implicit bias related issues, (stereotyping and attitude change disassociated from explicit beliefs and conscious processing), is expressed by Jonathan Evans:

> Dual process theories of social cognition emerged in the 1980's and developed in popularity to form the dominant paradigm for the past 20 years or more. Contemporary work particularly concerns the automatic and unconscious processing of social information in such domains as person perception, stereotyping, and attitude change and its apparent *dissociation from explicit beliefs and conscious processing*. The proposal of new accounts or at least new labels for Dual Processes in social cognition has reached near epidemic proportions … . (2008: 268)[112]

As the title of Evans' Paper suggests, there are related but different Dual Process theories of cognition depending on which domain is considered; (i) reasoning, (ii) judgment and decision making, (iii) social cognition; essentially how individuals construe the social world and the processes that underlie social judgement and behaviour.[113] Evans points

[112] I have removed Evans' detailed referencing and added my own emphasis.

[113] From UCL University guide (2018), MSc Degree, Social Cognition: Research and Applications.

out there are differences of emphasis between the different fields of research; in general, 'social cognition literature is less concerned with issues about cognitive architecture and evolution and more focused on issues concerning consciousness, free will, and the implications for moral and legal *responsibilities* of individuals' (added emphasis) (2008: 268). While noting Evans' comment about responsibility within the social domain, it will be recalled that within the judgment and decision-making domain, the idea of fast System One heuristic processes that cue default intuitive judgments endorsed by analytic System Two processes, also has implications for responsibility.

At this point it is useful to take stock and confirm the way forward within this chapter and the next. From consideration of the IAT, description of Dual Process and Dual System approaches to cognition naturally followed. S*ocial* cognition and the *unified* Dual System approach have been described and description of two particular Dual System models will follow shortly, together with further discussion concerning the possibility of overriding or moderating the influence of implicit bias. In the next chapter this theme will continue with description of a unified Dual System model of social cognition that will be used during critique of semicompatibilism in Chapter 6.[114]

While considering the possibility of overriding or moderating the influence of implicit bias related behaviour, (with implications for responsibility), two Dual System models will be described with reference to Bertram Gawronski and Laura A. Creighton's Paper, *Dual Process Theories* (2013). There is much to consider in this Paper, but attention will focus on Dual System theories/models that seek to describe the mechanisms that; (i) enable attitudes to guide behaviour, and (ii) may account, from a social cognition point of view, for only a moderate reduction in racial conflicts, against a background of declining negative evaluations of racial minority groups in public opinion polls.

[114] As mentioned in Footnote 107, the distinction between Dual Process and Dual System should always be kept in mind - there is the possibility of confusion as 'social cognition' seems to relate to a domain specific Dual *Process* rather than the more inclusive Dual System model. However, I believe that the complexity of social cognition is such that the description Dual *System* is considered appropriate.

Two models are described by Gawronski and Creighton under item (i). First, the *M*otivation and *O*pportunity as *De*terminants (MODE) model[115] developed by Russell Fazio. Attitude is considered as the mental association between an object and the agent's evaluation of that object. Assume a situation where association is strong, an immediate, good and established attitude is experienced towards a particular person when they are seen. In this case the spontaneous good attitude will guide behaviour without the individual necessarily being aware of the attitude's influence. Alternatively, if a strong attitude has not been previously established, an individual may examine specific attributes of a person (or object) and the situation, but this activity relies on motivation and opportunity (adequate time and cognitive resources) to engage in effortful information processing. If motivation and/or opportunity is low, automatically activated attitudes may guide behaviour based on spontaneous inference or interpretation of the situation. However, if both motivation and opportunity to engage in effortful processing are high, the impact of automatically activated attitudes on behaviour will be moderated and behaviour will be *subject to* consideration of the situation, including specific attributes of the person or object. In other words, there are two distinct processes that are guiding behaviour, one essentially based on spontaneous processes (System One) and the other on deliberative processes (System Two); the process (implicit or explicit social cognition) that will become *active* is determined or moderated by the agent's motivation *and* opportunity to engage in deliberative processing.

The second model, to be mentioned briefly, is the Dual Attitude Model[116], developed by Timothy D. Wilson, Samuel Lindsey and Tonya Y. Schooler. This model differs in the sense that

> previous approaches adopted a 'between - subjects approach', whereby different individuals with different kinds of attitudes are said to act differently. Our approach is 'within – subjects' in that the same individual can have both an

[115] See also Russell Fazio, *Multiple Processes by which Attitudes Guide Behavior: The MODE Model as an Integrative Framework* (1990), cited by Gawronski and Creighton.

[116] See Timothy D. Wilson, Samuel Lindsey and Tonya Y. Schooler, *A Model of Dual Attitudes* (Wilson, Lindsey, and Schooler 2000). See also the work of Anthony Greenwald, for example, *A Unified Theory of Implicit Attitudes, Stereotypes, Self-esteem, and Self-concept* (Greenwald et al. 2002).

> implicit *and* explicit attitude, which predict different kinds of behaviours. (2000: 121)

The notion of one individual with implicit *and* explicit attitudes that influence various kinds of behaviour is at once relevant to the subject matter of implicit bias, while keeping in mind that characterising or describing implicit bias as an 'attitude' is not universally accepted.

Turning to item (ii) above, Gawronski and Creighton's discussion of prejudice and stereotyping from a social cognition point of view begins with the observation of a moderate reduction in racial conflict when negative evaluations of racial minority groups in public opinion polls was in marked decline. It is believed the essential reason for this apparent inconsistency is change in the way racial prejudice is manifest, while clearly not abandoned. As an example of work inspired by this change, Gawronski and Creighton cite Patricia G Devine's influential Paper[117] *Stereotypes and Prejudice: Their Automatic and Controlled Components* (1989). The primary goal of Devine's three reported studies is to 'examine how stereotypes and personal beliefs are involved in responses toward stereotyped groups' (1989: 6). The central claim of this exceptionally interesting and essentially optimistic[118] Paper is that a distinction should be made between *knowledge* of a social stereotype and *belief* in the accuracy of that stereotype. This dissociation model assumes that when an agent encounters an object or person, automatic stereotype activation occurs, equally strong and inescapable for both high prejudice and low prejudice individuals. However, the two groups differ significantly in that low prejudice individuals can intervene, replacing automatically activated stereotypes with nonprejudicial beliefs (following Gawronski and Creighton 2013: 288). With time and

[117] Devine's Paper *Stereotypes and prejudice: Their automatic and controlled components* received the Scientific Impact Award from the respected and influential Society of Experimental Social Psychology. The award recognized that the Paper would have a lasting impact and had fundamentally altered and would inspire the future landscape of prejudice and stereotype research.

[118] See for example the following extract from the General Conclusions as evidence of an essential optimism: '(It has been argued by others that) … inconsistency sometimes observed between expressed attitudes and behaviors that are less consciously mediated is evidence that (all) White Americans are prejudiced against Blacks and that nonprejudiced responses are attempts at impression management. […] In the context of the present model in which automatic processes and controlled processes can be dissociated, I disagree fundamentally with this premise' (Devine 1989: 15).

intention, it is possible to change personal attitudes and beliefs, but sometimes effort may be lacking, and this can contribute to the often-observed inconsistency between expressed attitudes and seen behaviour. The claim that change is possible is worth reemphasising; the question is not if but how change is to be made.

To clarify the difference between the dissociation model and the MODE model; the dissociation model locates extrinsic social influences, that over a long-time feed into development of stereotypical models, at the level of automatic processes. The individual's authentic self is placed at the level of *controlled* processes that potentially mitigate behaviour originating from acquired stereotypes. The MODE model locates an individual's authentic self at the level of *automatic* processes and extrinsic social influences at the level of controlled processes. The moderate reduction in racial conflicts when negative evaluations of racial minority groups in public opinion polls was in decline is not mentioned within Devine's Paper specifically. One possibility, (a too obvious interpretation perhaps), is that in harmony with the dissociation model, where the individual's authentic self is placed at the level of *controlled* processes that potentially mitigate behaviour originating from acquired stereotypes, opinion polls give greater opportunity for reflection and evaluation of racial stereotypes. This would lead to a thoughtful assessment before expressing an anonymous opinion. Compare such anonymity with emotional and public conflict that 'flares up' almost spontaneously. It is possible to think carefully, calmly, rationally about what we say in an opinion poll, while in a conflict situation response is automatic, perhaps unconscious and it is this type of behaviour that is not influenced by increased conscious awareness and egalitarian ideas. Clearly, the MODE model is also suggested by the comment that opinion polls give greater opportunity for reflection. This matter clearly requires thorough analysis to arrive at any meaningful conclusions, but there is value in noting the potential usefulness of such models when investigating these issues.

Some general and concluding remarks on the Dual System model can be made with reference to Daniel Kahneman's Paper *A Perspective on Judgment and Choice, Mapping Bounded Rationality* (2003). The contribution of Daniel Kahneman and Amos Tversky towards the understanding of intuitive judgment and decision making has been immense. This Paper is based on the author's Nobel Prize lecture, delivered at Stockholm

University on the 8th of December 2002. The key claims made from a System One and Two point of view are as follows (2003):

> The (System One and Two) model suggests five ways in which a judgment or choice may be made:
> 1. An intuitive judgment or intention is initiated, and
> (a) Endorsed by System Two, or
> (b) Adjusted (insufficiently) for other features that are recognized as relevant, or
> (c) Corrected (sometimes overcorrected) for an explicitly recognized bias, or
> (d) Identified as violating a subjectively valid rule and blocked from overt expression.
>
> 2. No intuitive response comes to mind, and the judgment is computed by System Two.
>
> Casual observation suggests that Cases 1(a) and 1(b) are the most common and that Case 1(d) is exceedingly rare. This ordering reflects two major hypotheses about the role of intuition in judgment and choice. The first is that most behaviour is intuitive, skilled, unproblematic, and successful. The second is that behaviour is likely to be anchored in intuitive impressions and intentions even when it is not completely dominated by them.
>
> 3. Another testable[119] hypothesis is that *intuitive judgments that are suppressed by System Two still have detectable effects*, for example, in priming subsequent responses (added emphasis).
>
> 4. An intuitive judgment will be modified or overridden *if* System Two identifies it as biased. This argument is not circular because a great deal is known about the conditions under which corrections will or will not be made and because hypotheses about the role of System Two can be tested. In the context of an analysis of accessibility, the question of when intuitive judgments will be corrected is naturally rephrased: When will corrective thoughts be sufficiently accessible to intervene in the judgment?
>
> 5. The evaluation of stimuli as good or bad is a particularly important natural assessment. The evidence, both behavioural and neurophysiological is consistent with the idea that

[119] Daniel Kahneman's Paper *A Perspective on Judgment and Choice, Mapping Bounded Rationality* (2003) presents detailed description of test methods, substantial analysis and presentation of resulting data.

the assessment of whether objects are good (and should be approached) or bad (and should be avoided) is conducted quickly and efficiently by specialized neural circuitry.

Unfortunately, it is not possible to look in detail at the above claims and conclusions from Daniel Kahneman's Paper, however, a *vital* point must be noted, that an essentially optimistic outlook is presented in the sense of the executive role of System Two managing and monitoring behaviour. This appears to be at the heart of the matter; if the source of implicit bias influenced behaviour is located within System One processes then clearly the *possibility* of System Two control (even degrees of control) of the 'manifestations of System One activity' (Evans and Frankish 2012: 28) is crucially important, not least because of the implications for responsibility. That intuitive judgment may be modified, moderated or overridden *if* System Two identifies it as biased strongly suggests control and responsibility. One of the widely assumed characteristics of implicit bias, mentioned many times, is absence from conscious *awareness* in thought and action and so unavailable for scrutiny and intervention by System Two processes. The above claims from Daniel Kahneman's Paper and much of what has been discussed in this chapter so far push back against this pervasive view and show the possibility of System Two control and responsibility. The subtle nature of implicit bias will be examined further, and in Chapter 5 control will be seen to be possible, for example, from the perspective of control asserted by the true self.[120] Given that it can be argued successfully that *control* in some significant sense *is* possible, this suggests that implicit bias may not be threatening to certain positions, such as semicompatibilism, within the free will debate. There are several aspects of control; control over acquisition of implicit biases, control in the sense of eliminating biases from our cognitive processes and control of behaviour that has implicit bias as part of its determination. Consideration of freedom should include all these areas, at least, where implicit bias is active; it is however control of influenced behaviour that will be the main but not exclusive area of interest.

[120] See also *Systems and Levels: Dual-System Theories and the Personal—Subpersonal Distinction* (Frankish 2012). The possibility of 'true self' control is a popular claim, suggested indirectly, for example, in the well-respected *Implicit Bias Review*. In this particular quotation, the possibility of controlling implicit bias influenced actions via our 'true intentions' is thought to be problematic; 'We act on our implicit biases without awareness, thus, they can undermine our *true* intentions' (Staats et al. 2016: 14).

Dual System models attract criticism,[121] usually based on the difficulties experienced when trying to decide exactly what sort of empirical evidence could disprove such claims and the lack of empirically testable predictions. However, as Bertram Gawronski and Laura A. Creighton conclude, '[…] despite such criticism of Dual Process and Dual System theorizing, it seems highly unlikely that this influence will dissipate in the near future' (2013: 308).

The Nature and Freedom of Intuitive Thought and Decision Making (Jarvilehto 2015), particularly Chapter Two, has much to say about Dual Systems. It is however the notion of a third system and its relevance to implicit bias, in terms of the crucial role of the environment in bias formation and ongoing influence, that suggests mention here. To understand behaviour, the environment in which that behaviour takes place must be considered. This seems entirely obvious, but when framed in terms of System One, Two and Three, where System Three is responsible for generating the context for action, an interesting model of cognition within an environment can be developed, as reproduced in Fig 4.6 (Jarvilehto 2015: 51). The three systems form a nested structure with the conscious agent at the centre. The agent acquires information from the non-conscious System One, which in turn is constantly influenced by events and changes in the environment, System Three. There is no representation of environmental input directly into System Two processes, but this is surely the case. This model raises several such questions, but there is value in mentioning such a descriptive model of cognitive processes and environment, particularly considering the huge importance of environmental factors in any discussion of implicit bias.

[121] See for example, David Sorensen's Thesis, *The Unity of Higher Cognition: The Case Against Dual Process Theory* (2016). For a general defence of Dual Process and Dual System models against five lines of critique see Evans and Stanovich *Dual-Process Theories of Higher Cognition: Advancing the Debate* (2013).

This Figure is available from *The Nature and Freedom of Intuitive Thought and Decision Making* (Jarvilehto 2015: 51).

Fig. 4.6 Three Nested Systems: System One Non-conscious, System Two Conscious and System Three Environment. (Jarvilehto 2015: 51)

System One is shown in Fig 4.6 differentiated according to ontogenetic and phylogenetic processes, where ontogenetic processes are acquired through experience and practice, for example, skills and decision-making heuristics. Phylogenetic processes have their origin in, what Jarvilehto calls, the biologically evolved environment, for example protective parental behaviour and fight or flight. System Three is shown differentiated into the culturally evolved and biologically evolved environments and System Two shows a differentiation between the algorithmic mind, (slow thinking and computation) and the reflective mind.[122] Although clearly very brief, mention of a third system representing the environment highlights and formalises in some sense the obviously connectedness of environment, modes of cognition and the agent. Such connections will be explored in much greater depth in the following section where the focus will be the nature of implicit bias.[123]

[122] For an appropriate treatment of System Three and associated ideas see Keith Stanovich*, Distinguishing the Reflective, Algorithmic, and Autonomous Minds: Is it time for a tri-process theory?* (Stanovich 2009: 55).

[123] See also Clark's description of ecological control *Soft Selves and Ecological Control* (2006) in the next Chapter, Implicit Bias and Control, and *An Introduction to Implicit Bias* (Beeghly and Madva 2020) that includes several Papers discussing the crucial role of the environment in bias formation and ongoing influence.

Much ground has been covered in this section. It is important to make a summary and clarify the relevance to what will follow. The IAT, (controversially), is a rigorous method of testing for the presence and degree of implicit bias based on measured time differences between more or less automatic responses to a carefully developed set of images that have, for example, racial and evaluative content. This led naturally to discussion of Dual Process and Dual System models of cognition. Dual System models offer a 'rich descriptive paradigm' of implicit bias, learning and cognitive structures. How evolutionary and heuristic paradigms were reconciled by introducing a *unified* Dual System model was described. Social cognition was mentioned, including the MODE and dissociation model. Implicit bias and Dual Process/System models of cognition are significantly entwined, as suggested by Devine's important Paper *Stereotypes and Prejudice: Their Automatic and Controlled Components* (1989). The purpose of this section has been to give context and prepare for the introduction of the model of implicit bias to be described shortly; to give necessary background in a comparable way to the opening chapters of Part I.

4.3 The Origin of Implicit Bias

It is sobering to reflect on one person's experience of racial bias and reconfirm the real-world implications of this phenomena. A poignant memory is described by Laurie A. Rudman at the beginning of her Paper *Social Justice in Our Minds, Homes, and Society: The Nature, Causes, and Consequences of Implicit Bias* (2004):

> In 1964, when I was 10 years old, my dad drove off with my older sister, Carol. When he returned, my parents announced she would not be coming back. She wanted to marry Lenny — a young African American she had met in college. I couldn't understand what the problem was. The one time he came to our house, I sat on his lap and was thrilled with the way he laughed at my jokes. He was generous and kind and, in fact, years later, he would head the United Way in Chicago[124] My father had taken Carol to Northeast Minneapolis and told her to

[124] The United Way of Metropolitan Chicago is a non-profit organization and a branch of the United Way of America (now United Way Worldwide). The United Way of Metropolitan Chicago serves the city of Chicago and its surrounding suburbs, allocating funding to other charitable organizations, especially those that provide needed healthcare, education and income services to underserved communities. (From Wikipedia *United Way of Metropolitan Chicago* 2017).

> make a choice. It was either 'those people' or her family. My sister got out of the car. (2004: 129)

What *is it* that has the power to cause[125] such extreme behaviour and upset within a family; how can implicit bias be understood?[126] The widely perceived attributes of implicit bias have been mentioned: pervasive, unconscious and involuntary, where bias 'denotes a displacement of people's responses along a continuum of possible judgments' (Greenwald and Krieger 2006: 950) and may tend towards a favourable or unfavourable assessment. How such bias is formed is controversial but obviously important not least in terms of developing and implementing prevention and mitigation strategies.

The main question raised at the beginning of this chapter concerned the nature of implicit bias.[127] It is the widely held view of implicit bias as nonconscious and source of involuntary behaviour that first suggests possible threats to free will, control and responsibility.[128] Expressed in terms of the semicompatibilist model, can behaviour influenced by implicit bias issue from the agent's own reason-responsive mechanism when the agent is not consciously aware of possessing or being influenced by implicit bias? The attributes of nonconscious and involuntary will be considered, within a larger group[129] of implicit bias related issues, ideas and concepts.

To be explored in this section: Attitude and belief, association and stereotype, and the propositional structure of implicit bias. Responsiveness to reason is particularly

[125] It should be noted that describing implicit bias as 'causing' behaviour is not universally accepted. See for example *Implicit Attitudes and the Ability Argument* (Buckwalter 2018).

[126] For powerful description of the psychological damage caused by colonialism and racism see particularly *Black Skin, White Mask* (Fanon 2007).

[127] Implicit bias related research has been recently subject to various forms of criticism. Three leading academics and researchers in the field, Michael Brownstein, Alex Madva and Bertram Gawronski respond in their Paper *Understanding Implicit Bias: Putting the Criticism into Perspective* (2019).

[128] See for example *Consciousness, Implicit Attitudes and Moral Responsibility* (Levy 2014a), *Implicit Attitudes and the Ability Argument* (Buckwalter 2018) and *Implicit Bias, Responsibility, and Moral Ecology* (Vargas 2017).

[129] There is diversity in the way the actual term Implicit Bias is used. See Jules Holroyd and Joseph Sweetman's discussion of 'the important functional differences between phenomena identified as instances of implicit bias', *The Heterogeneity of Implicit Bias* (2016: 80).

important in assessing whether implicit bias is subject to the sort of control (guidance control) sufficient for responsibility within the semicompatibilist model.

Social psychologists define an *attitude* as an evaluative disposition, the tendency to like or dislike, or to act favourably or unfavourably toward someone or something. Explicit expression of attitudes happens often, either verbally, by action, or both means of expression working together. A social *stereotype* is a mental association between a social group or category and a trait. It may be the case that a group or category display a certain characteristic, (basketball players display physical stamina), however it is not always the case that every member of a group must display a certain characteristic for a stereotype to be formed. If a small minority (10-15%) of drivers over seventy-five years old drive twenty miles per hour under the speed limit then it may come to serve as a default assumption that any elderly person is likely to drive slowly (following closely Greenwald and Krieger 2006: 949). Groups having *no* statistically meaningful tendency towards a particular trait may still be subject to stereotype association, either favourable or unfavourable, caused by untruthful representation by, for example, the media and/or distorted or manipulated historical narrative. A stereotype is the association of a particular trait with a particular group. An attitude is the association of a particular evaluation, such as good, honest, untrustworthy, with a particular group. *Implicit biases* are discriminatory biases based on implicit attitudes or implicit stereotypes (following Greenwald and Krieger 2006: 951).[130]

From the beginning of this chapter association has been the central explanatory paradigm of implicit bias, discussion began with the IAT, a test predicated, (together with the notion of differential processing speeds), on a model of association between negative attitude and concept. Eric Mandelbaum notes the central and pervasive role of

[130] An alternative account is given by Kang, who uses the term 'schema' to describe a wide range of information about the attributes of a concept under one heading (2005: 1498). A schema is a prototype or template for a class of objects, 'a mental shortcut that allows quick assignment of objects, processes or people into categories. For example, people may be placed into categories based on traits such as age, race, gender, and the like' (Kang cited in Staats 2013: 11). When a category has been assigned to a person *all* the attributes associated with that category become associated with that person. It is not difficult to see how this idea works for racial schemas. Kang suggests the law, culture and society generally are the source of racial categories into which individual human being are mapped. 'Once a person is assigned to a racial category, implicit and explicit racial meanings associated with that category are triggered' (2005: 1499).

association, claiming 'the overwhelming majority of those who theorize about implicit biases posit that these biases are caused by some sort of association' (2016: 1), where 'saying that two concepts are associated amounts to saying that there is a reliable, psychologically basic causal relation that holds between them' (2017). On this view, implicit biases have an associative structure, enter into associative transitions and do not enter into logical ones, (following Mandelbaum 2016: 6). Mandelbaum continues, claiming that associative structures may *only* be broken by repetitive presentation of one of the objects between which a relation is said to hold without the other or by the counter-conditioning effect of a repeated valence change relating to one of the objects. If rational argument *is* seen to have an impact on implicit attitudes, then implicit attitudes cannot have a completely associative structure because associative structures simply do not respond to logical arguments. Having the right position on this matter is obviously important when developing mitigation plans, from a personal and from wider perspectives such as education, employment, healthcare and justice.

Mandelbaum *challenges* the ubiquitous association model, beginning with a description of an alternative explanatory paradigm, the structured belief hypothesis. As the name suggests, implicit biases are sustained by unconscious beliefs, *not* associations, beliefs that are 'propositionally structured mental representations that we bear the belief relation to' (2016: 7). This clearly runs counter to conventional, pervasive, purely System One based explanations of implicit bias that deny the possibility of unconscious structured representations entering logical relations with one another. A key claim here is structured beliefs can be *reason-responsive* and if implicit biases are sustained by structured beliefs, they will also be reason-responsive. If true, this has especially important implications for implicit bias mitigation, responsibility and ultimately discussion of personal freedom. Having suggested an alternative to pervasive association-based models of implicit bias, Mandelbaum challenges associationism from several perspectives. I will describe one perspective, beginning with an outline of the essential argument, as mentioned above.

The structure of Mandelbaum's argument is straight forward, beginning with a statement of the associationist hypothesis that implicit biases have associative structure, enter into associative transitions and do *not* enter into logical ones, (Associative Implicit Bias, AIB). If this hypothesis is true, then reliable and successful intervention can only take the form of an 'extinction process', i.e., multiple presentation of just one (the Conditioned Stimulus) of the associated elements, or counterconditioning, where the CS is presented together with a 'reinforcer' that has an opposite valence to the one the CS currently has. *If* intervention by rational argumentation (or *any* logical or evidential intervention) is found to reliably counteract implicit bias and 'if these interventions are not reducible to extinction or counterconditioning, then we have *evidence* that the structure of implicit bias is not, after all, underwritten by associations' (Mandelbaum 2016: 9).

Mandelbaum presents such evidence from various perspectives to support the claim that implicit attitudes have more structure than mere associations, believing that 'we have a handle on what structure that is: mental representations with propositional structure that function as unconscious beliefs' (2016: 18). I will briefly describe the first perspective.

Mandelbaum describes as 'a venerable social psychological hypothesis' (2016: 10) the well-known idea that the enemy of our enemy is our friend, discussed formally by Fritz Heider as Balance Theory, *The Psychology of Interpersonal Relations* (Heider 1958). See also Eva Walther, *Guilty by Mere Association: Evaluative Conditioning and the Spreading Attitude Effect* (2002).

Mandelbaum argues if mental transitions are merely associative, (*not* inferential), evidence of normal second-order conditioning effects would be expected, contra to what is predicted by Balance Theory. Mandelbaum's example clarifies the point. The *associationist* predicts that if I have a negative association with Assad who I know has a negative association with Szymborska, this should lead, on the associationist account (and contra to Balance Theory), to a *negative* association with Szymborska, for Szymborska has been paired with two negative stimuli. *If* evidence is available that shows a subject responding in a way that supports Balance Theory, (the enemy of our enemy is our friend), then such evidence gives confidence to the claim that implicit attitudes are *not* purely associative processes but 'have some sort of logic operating over them' (2016: 10). Expressed in terms of negatives and positives, such evidence would show two negatives making a positive; 'if someone has been mean to you, you will tend to like people who are mean to your antagonizer' (2016: 10).

Gawronski and colleagues *Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information* present from their first experiment, empirical evidence that *does* show two negatively valenced implicit attitudes leading to a positive evaluation (2005: 619-622). Gawronski's method and results from Experiment 1 are summarised below:

- Introduction of subjects to a photo of an unfamiliar individual (CS1).
- CS1 paired with statements either consistently positive or consistently negative, thus conditioning the subjects to respond to the CS1 with the designated evaluation.
- CS2 was introduced and subjects were told whether the CS1 liked *or* disliked the CS2.
- The subjects' implicit attitudes toward both the CS1 and the CS2 were assessed.
- The procedure was then replicated for five other novel CS1 and CS2 pairs.

Gawronski's results support Heider's Balance Theory predictions and, importantly, do not support the predictions of the AIB model; 'A negatively valenced CS1 who disliked a CS2 caused the subjects to *like* the CS2. In other words, if you were originally taught that a person was bad and subsequently learned that this person dislikes another person, you then would like that second person' (Mandelbaum 2016: 11). On Mandelbaum's view, two such negatives making a positive shows a propositional, and not an associative, process is present and active.

In summary, data and conclusions from Gawronski's first experiment (2005: 619-622) are central to Mandelbaum's argument that implicit attitudes have neither associative structure nor enter into associative transitions (contra AIB). Further, on this view, the data demands propositional processes and inferential structures, this lends support to a Structured Belief-type view (following Mandelbaum 2016: 12).

As mentioned, the claim of the reason-responsiveness of implicit attitudes is of considerable importance at least in terms of mitigation of implicit bias. Such critical review of associationism is extended by Mandelbaum towards Dual System models within the discipline of social psychology. It will be recalled that typically System One processes are described as fast, *automatic*, intuitive, non-rational, unconscious and *associative*. The essential criticism is that it is wrong to infer from any System One property to another. If a process is automatic, it is not necessarily non-rational or heuristic. If it is unconscious, then it is not necessarily associative. The theme of

Mandelbaum's argument, *contra* to most Dual Process theories, is that propositional structures *do inhabit* the System One domain and System Two processes are present and active within the supposedly exclusive world of System One. Deconstruction of the pervasive System One and Two model is completed by Mandelbaum claiming that 'propositional processes and structures not only affect unconscious states, but the propositional structures can *be* unconscious states and their corresponding logical processes can operate unconsciously' (2016: 19).

Structured belief as a foundation for implicit bias is not without problems. While Dual Process/System theories can explain how implicit attitudes and explicit attitudes often *differ*, it is not clear how structured belief can accommodate this familiar claim. However, an explanation is available, based on the idea of fragmented beliefs. Fragmented beliefs, as the name suggests, are causally isolated from each other, hence able to exist simultaneously within one mind. There is no single consistent set of beliefs, fragmented beliefs may be in some cases contradictory. An alternative perspective is the introduction of a situation-specific dimension within the description of an attitude. Irene V. Blair (2002: 256) describes this idea, quoting from Abraham Tesser's article *Self-Generated Attitude Change*:

> An attitude at a particular point in time is the result of a constructive process ... And, there is not a single attitude toward an object but, rather any number of attitudes depending on the number of schemas available for thinking about the objects. (Tesser 1978: 297)

Some key terms within the implicit bias debate have been described, i.e., attitude, belief, association and stereotype. An important alternative to the widespread associationist model has also been described based on structured beliefs, developed considering the claimed ineffectiveness of counterconditioning to change implicit attitudes. Mandelbaum's structured belief hypothesis offers the particularly important possibility of mitigating implicit bias by rational intervention and argument, as structured beliefs, unlike associations, are essentially reason-responsive, a vital concept within semicompatibilism.

4.4 Summary

Recalling the question raised at the beginning of this chapter; what is the nature of implicit bias? After introducing implicit bias, the IAT was described, the most well-known testing model for implicit bias, more important, it was noted that the test is based on the pervasive view that association is at the heart of the implicit bias process. This led to description of Dual Process and Dual System models of cognition whereby differential response speeds, essential measurements within the IAT, are explained in terms of implicit bias as a System One process. If implicit bias is essentially a System One nonconscious associative process, then control or override by System Two conscious rationality seems problematic. Eric Mandelbaum challenges the associative model using experimental data and a structured belief hypothesis where implicit biases are sustained by unconscious beliefs that *are* propositionally structured mental representations that *can* be reason-responsive.[131] Clearly, the notion of control, (for example by conscious System Two processes), that originates from rational responsiveness to argument from others, and within the agents' own internal deliberation, is surely vital if implicit biases are to be managed at a personal or institutional level. Such rational responsiveness described by Eric Mandelbaum is clearly suggestive of the notion of reasons-responsiveness within John Martin Fischer's semicompatibilism.

If the existence of implicit bias and its influence on behaviour are not easy, in some cases perhaps impossible, to discern consciously, then it is unsurprising that behaviour as a manifestation of implicit bias is also *difficult* to control and mitigate at a personal and corporate level.[132] However, there is clearly 'a world of difference between

[131] See Keith Frankish *Systems and Levels: Dual-System Theories and the Personal — Subpersonal Distinction* (2012). On this account, at the subpersonal level there exists hybrid systems having some System One properties and some System Two properties. *Personal* reasoning constitutes a distinct level of mental activity, which can be clearly distinguished from the lower, subpersonal one. Also, Frankish *Playing Double - Implicit Bias, Dual Levels, and Self-Control* (2016). This wide-ranging Paper concludes that to suppress implicit bias it is not sufficient to have explicit unbiased belief and a desire to act accordingly. In addition, the agent must have a strong *implicit* desire to make those explicit propositional attitudes effective in reasoning and action; strength of will (2016: 42).

[132] For example, see *Why Diversity Programs Fail* from the July/August 2016 issue of the Harvard Business Review.

being difficult and impossible to control' (Madva 2018: 62). It is argued within this Thesis that empirical evidence supports the former claim, that controlling implicit discrimination can be difficult, taxing, demanding but is possible (following Madva 2018: 62). In the next chapter, an interactive Dual System, developed by Deutsch and Strack, *Building Blocks of Social Behaviour* (2010), will be described that shows control of implicit discrimination is possible.

I agree with Levy that an agent is morally responsible for an attitude or an action iff it is appropriately attributable to the agent (Levy 2017: 4), where the condition 'appropriately attributable' is satisfied in cases where the agent *should* be aware of their implicit bias.[133] [134] Mandelbaum's claim that 'propositional processes and structures not only affect unconscious states, but the propositional structures can *be* unconscious states and their corresponding logical processes can operate unconsciously' (2016: 19) is clearly important not only because such a view presents a robust *alternative* to the widely held associationist/Dual System/Dual Process view, but also offers positive grounds for an understanding of implicit bias as reason responsive and subject to control and responsibility.

This chapter began with a brief and uncritical look at the IAT. As previously noted, this set the scene for discussion of implicit bias within the context of the Dual System model of cognition. It is beyond the scope of this Thesis to appraise apparently conflicting empirical data[135] concerning reason-responsiveness of implicit attitudes or propositions, however, at the time of writing (June 2018) it is generally accepted that

[133] For example, a hiring committee has a responsibility to be at least minimally aware of fundamental advances in bias and prejudice scholarship. Responsibility is characterised as a 'function of the external context and wider social circumstances' inhabited by the agent. Ignorance of implicit bias is not necessarily an exculpating condition of responsibility for behaviour where implicit bias is the source (Washington & Kelly 2016: 24).

[134] It is recognised that 'should be aware' is problematic and will be mentioned again during later discussion of awareness and responsibility within earlier cultures that supported behaviour now considered to be wrong.

[135] As an example of conflicting positions, refer to *The Malleability of Automatic Stereotypes and Prejudice* (Blair 2002), where the idea that behaviour influenced by implicit bias is automatic hence inevitable and nearly impossible to avoid is challenged by the conflicting position ' … that the perceiver's focus of attention *can* influence the automatic operation of stereotypes and prejudice, as well as more controlled processes' (added emphasis) (Blair 2002: 251).

implicit bias *is* malleable (Cheryl Staats 2013: 53), although this claim is controversial and complex, see for example, *Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences* (Gregg A, Banaji M, Seibt B 2006). Further, there is no clear winner in terms of a mitigation strategy, and medium to long term effectiveness continues to be a problematic.[136]

Having looked at the associationist model of implicit bias and a radically different perspective developed by Mandelbaum, the next chapter looks at implicit bias and control; a model of implicit cognition to be used in Part III will be developed that accepts a greater or lesser degree of conscious awareness and includes a reflective component, facilitating control of the behavioural expression of implicit bias and so responsible behaviour. This is obviously important because absence of *control* strongly suggests an absence of agent responsibility for issuing behaviour. Further, moving towards a clear position on control and responsibility for implicit bias related behaviour is necessary preparation for addressing the main question, does implicit bias threaten the semicompatibilist position on free will and responsibility.

[136] See *State of the Science: Implicit Bias Review 2013*, particularly Chapter 8, Debiasing (Cheryl Staats 2013: 53).

Chapter 5

Implicit Bias and Control

> *Implicit biases aren't just coloring our thoughts, perceptions, and actions from behind the locked door of the unconscious but are themselves palpably present (or at least accessible) to awareness.*[137]
>
> Alex Madva

5.0 Introduction

At the beginning of the previous chapter, the apparent threat to free will posed by implicit bias was simply expressed in terms of the influence of implicit bias on behaviour without agent awareness, removing the possibility of agent intervention or conscious choice and so removing the possibility to act in available alternative ways. In other words, an absence of *control* and hence, it is commonly claimed, an absence of responsibility for issuing actions. On this view, an agent is, using Madva's (2018: 66) term, 'exonerated' from responsibility by absence of awareness of their implicit biases or implicit bias influenced actions. In the absence of conscious awareness, responding to reasons and making rational alternative choices seems impossible. This can be framed in terms of a simple Dual Process model whereby behaviour issuing from implicit bias is the outcome of automatic System One processes, not subject to System Two conscious control or mediation. However, it will be recalled that Eric Mandelbaum developed a *challenge* to the often-discussed associative and Dual Process models of implicit bias using experimental data and a structured belief hypothesis where implicit biases are sustained by unconscious beliefs that are propositionally structured mental representations that *can* be reason responsive (page 107).

[137] *Implicit Bias* (Madva 2020: 387). See also Madva *Implicit Bias, Moods, and Moral Responsibility* (2018: 70).

In the Introduction to the important Paper *Attributionism and Moral Responsibility for Implicit Bias* (2016a),[138] Brownstein clearly expresses the issue of implicit bias, awareness, control and responsibility, and encouragingly supports implicit bias as a test case for theories of moral responsibility:

> What we know about implicit attitudes suggests they do not easily fit into traditional philosophical approaches to theorizing about moral responsibility. For example, in the empirical literature, 'implicit' typically means outside of conscious awareness or control. Philosophers who think that moral responsibility hinges on agentive control over, or awareness of, one's attitudes may take this usage to suggest that people are not responsible for their implicit biases. […] *Implicit bias is therefore a good test case for theories of moral responsibility* that aim to accommodate the messy reality revealed by contemporary sciences of the mind. (added emphasis Brownstein 2016a: 766)

On Brownstein's view, people *are* often aware of the content of their implicit attitudes but often unaware of the effects their implicit attitudes have on their behaviour. Interestingly, Brownstein differentiates implicit from explicit attitudes by virtue of their insensitivity to what is explicitly taken to be true or good (2016a: 770).

At the end of the last chapter it was stated, rather vaguely, that behaviour as a manifestation of implicit bias is difficult to control and mitigate at a personal and corporate level. Given such a range of positions, opinions and arguments, the aim now is to look deeper and consolidate what has been said so far, working towards a final unambiguous position on implicit bias in terms of control of issuing behaviour.

This chapter considers essentially two approaches to implicit bias, control and responsibility. The first is based around Wesley Buckwalter's Paper *Implicit Attitudes and the Ability Argument* (2018), followed by Holroyd and Kelly *Implicit Bias, Character, and Control* (2016). Buckwalter's Paper gives structure to the first investigation; the 'ability argument' critiqued by Buckwalter provides a convenient and more formal expression of the simply expressed popular claim that an agent is *not* responsible for actions that issue from implicit bias. This is important because ultimately the question to be answered

[138] In this Paper Brownstein argues 'that agents are morally responsible for actions that reflect upon what they *care* about, in the sense that these actions open them to being evaluated as moral agents. […] in paradigmatic cases, behavioural expression of implicit biases reflect upon agents' cares' (2016a) and so is legitimately subject to moral evaluation.

is, does implicit bias threaten semicompatibilism in the particularly important area of responsibility? *If* argument and empirical data convincingly conclude that an agent *is* responsible for behaviour issuing from implicit bias, and later (Chapter 6) such behaviour is shown *not* to be subject to guidance control and agent responsibility, as defined by the semicompatibilist model, then semicompatibilism is threatened in the sense that its response to control of implicit bias influenced behaviour appears to be incorrect. Under such circumstances something is wrong; either the characterisation of implicit bias or the semicompatibilist model, or perhaps both, require revision. It is semicompatibilism that is under investigation, so initially the model of implicit bias will be held fixed, unless convincing evidence emerges that revision is necessary. So, the aim within the next part of this chapter is to finalise the characterisation of implicit bias to be taken forward.

5.1 Implicit Bias and Control

The question is, are agents morally responsible for behavioural expression of their implicit attitudes and biases? Buckwalter's Paper *Implicit Attitudes and the Ability Argument* (2018) is essentially a critique of a particular response to that question i.e., the claim that an agent is *not* morally responsible for actions that issue from implicit bias. The claim that lack of control of implicit attitudes and behaviour entails an absence of responsibility for such behaviour. Buckwalter expresses the claim to be critiqued formally using the 'ability argument' (2018: 4):

1. S does *not* have the ability to control implicit attitude p
2. Implicit attitude p causes action Ø, therefore, S does *not* have the ability to control action Ø (by the Transfer of Responsibility Principle)
3. If S is morally responsible for Ø, then S has the ability to control Ø
4. Therefore, S *cannot* be morally responsible for Ø

Contrast the above with Holroyd and Kelly (2016), to be considered later;

1. If an agent has the relevant control and responsibility for implicit bias, then implicit bias reflects an agent's character.
2. Individuals *do* have relevant control/responsibility for implicit bias.

3. Therefore, implicit bias reflects an agent's character, (and *can* legitimately be morally appraised).

Considering the key concepts of control, causation and responsibility, Buckwalter's Paper tries to refute the ability argument; that agents are *not* responsible for actions issuing from implicit bias. Buckwalter attacks the argument on all fronts, claiming all the premises are incorrect, and even if they were correct the argument is invalid. At the outset Buckwalter flags an issue with the ability argument because it invokes the questionable Principle of Transfer of Powerlessness (see page 41), whereby lack of control by S of p is transferred to lack of control of Ø. While initially plausible, this principle has issues that will be discussed later, for the moment I will look at the premises of the argument.[139] [140]

Consider Premise 1 and 2, both involving the notion of control; that S does not have the ability to *control* implicit attitude p and S does not have the ability to *control* action Ø caused by implicit bias.

Buckwalter claims with respect to the first premise ' … current evidence appears to point in the opposite direction' (2018: 5) and the substantial Paper *Meta-Analysis of Procedures to Change Implicit Measures* (Forscher et al. 2019) is cited as supporting this claim, (together with Papers by other leading authorities in the field, for example, Blair, Dasgupta, Greenwald and Frankish). Forscher's recent (2019) eighty-four-page analysis of over eighty thousand participants across three hundred articles published during the last twenty years concludes 'implicit measures[141] *can be changed,* but there is little evidence that changes in implicit measures translated into changes in explicit measures and

[139] See Fischer and Ravizza (1998: 18), Ravizza (1993) and Carlson (2003) for detailed discussion of the Principle of Transfer of Responsibility.

[140] See also *An Argument for Incompatibilism* (van Inwagen 2003).

[141] Implicit tasks assess associations through behavior that does not require deliberate retrieval of the target association. Explicit tasks assess associations through behavior that requires deliberate retrieval (e.g., answers to a questionnaire). *Tasks* are procedures designed to generate behavioural responses for data analysis. Tasks are distinguished from *measures*, which are defined as the *outcome* of a data-analytic technique applied to behavioural responses. On an implicit task, comparisons between responses that result from pairings between one set of concepts relative to responses from a different pairing is referred to as an *implicit measure* of response bias. Similar comparisons on an explicit task are referred to as an *explicit measure* of response bias (Forscher et al. 2019).

behavior, and we observed limitations in the evidence base for implicit malleability and change' (Forscher et al. 2019: 44). So, very measured support for the possibility of change of implicit attitudes. Buckwalter makes a much more positive interpretation, saying 'the principle finding of this research was to confirm that implicit attitudes *can* change and to identify which procedures were effective at changing them over others …' (2019: 5). Many examples of implicit bias mitigating strategies and studies are given by Buckwalter claimed to support the view that implicit biases *are* changeable. At this point, it seems reasonable to agree there is some measured support for the view that implicit biases are changeable, however, changeable and *controllable* (rejected by the first premise) are clearly not simply interchangeable terms.[142]

The change – control issue is not addressed, and Buckwalter continues, describing indirect responsibility, where it is reasonable to expect an agent to deliberately take earlier actions to be able to act well in the future, for example, a doctor keeping up with current research to ensure they act in the best possible interests of their patient later, (a reasonable assumption by the patient). Direct responsibility is also described, where control, responsibility and action occur at the same time. With respect to the first premise of the ability argument, (S does not have the ability to control implicit attitude p), from his reading of Forscher's *Meta-Analysis of Procedures to Change Implicit Measures* (2019), Buckwalter believes there is insufficient evidence available to counter *his* claim that such indirect and direct control (and so responsibility for implicit bias related behaviour) *is* possible (2018: 4). In other words, Buckwalter claims the first premise of the ability argument is not supported by sufficient evidence or good argument; there is no meaningful counter available to the claim that control and responsibility for implicit bias (and related behaviour) *is* possible.[143]

[142] When considering implicit bias and change, the issue of awareness is never far away - without awareness of bias how is change, or control, of bias possible? Because of its importance, awareness is discussed at several points within this Thesis.

[143] John Martin Fischer offers, as part of his guidance control model, the notion of 'tracing' whereby moral responsibility for an act at T requires the actual operation of a reasons-responsive mechanism at T *or some suitable earlier time* (1998: 50). For a detailed discussion of tracing see *The Place of the Trace: Negligence and Responsibility* (Murray 2019) and *The Truth about Tracing* (Fischer and Tognazzini 2009). This has clear implications for implicit bias related behaviour and is explored in Part III.

I will return to Buckwalter and conclude discussion of the second premise of the 'ability argument'[144] but first, further examination of change in terms of implicit measures and their translation into change in explicit measures and behaviour, and introduction of Deutsch and Strack's Model of mechanisms that may mediate the influence of implicit social cognition (attitudes, stereotypes, self) on behaviour.

The Deutsch and Strack Model

As the title of this chapter confirms, *control* of implicit attitudes and their influence on behaviour is clearly a central concern, and I will approach it with further observations on Forscher[145] and *change*. This is especially relevant, as considering how changes in implicit measures translate into changes in explicit measures and behavior will suggest a model developed by Deutsch and Strack (2010) that will be used extensively in Part III. This model describes the mechanisms that mediate the influence of implicit social cognition (attitudes, stereotypes, self) on behaviour. Sometime will be taken to describe this essential model and the control and responsibility implications for agents' behaviour expressed by their implicit biases. Forscher says that 'the results of the current meta-analysis do not lend themselves to a single interpretation' (2019: 45) and suggests possible explanations for implicit measures' relationship with explicit measures and behaviour.[146] However, Forscher concludes that the results of his analysis present a theoretical and empirical puzzle (2019: 44) and further work is believed to be necessary to understand the relationship between changes in implicit measures and changes in explicit measures and/or behaviour. The meta-analysis is ultimately inconclusive regarding whether implicit measures can be changed, and after repeated careful reading of Forscher's Paper it remains unclear *how* implicit bias active within a System One

[144] Discussion of Buckwalter and the second premise of the 'ability argument' continues page 125.

[145] Authors contributed equally to this manuscript and the order of the names decided by coin flip; Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, Charles R. Ebersole, Michelle Herman, Patricia G. Devine and Brian A. Nosek.

[146] The fourth possible explanation is particularly interesting, the possibility that automatically retrieved associations are causally inert. This is mentioned because if automatically retrieved associations *are* causally inert then basic assumptions of current research in implicit social cognition would have to be completely revised. There is, of course, considerable support *for* the position that implicit biases *are* causally linked to discriminatory outcomes as previously described at some length.

domain *could* affect explicit System Two deliberative behaviour. It is unclear how 'changes in implicit measures mediate changes in explicit measures and behavior' (2019: 38) is to be understood.

Within their discussion of associative versus propositional processes Deutsch and Strack (2010: 65) present an *explanation* of the *interaction* of System One and System Two processes, an explanation that will form the basis of the chosen model of mechanisms that mediate the influence of implicit social cognition (attitudes, stereotypes) on behaviour; a model that includes the possibility of control of behaviour. For Deutsch and Strack, associations have no claim to truth, no factual link necessarily exists between the associated elements, being formed 'on the bases of temporal contiguity and frequency of pairing' (2010: 65) within an impulsive system, (System One). By contrast, propositional representations and assessments of truth function within a reflective system, (System Two), taking some working 'data' from associative memory. Indirect measures such as the IAT address processes within the impulsive system and direct measures address processes within the reflective system. These are familiar ideas from Chapter 4, *but* Deutsch and Strack continue, claiming there is evidence that such clean demarcation is not the case; the two systems *interact*, impacting each other as described, for example, by Strack and Deutsch's Reflective-Impulsive Model (2004) and Gawronski and Bodenhausen's Associative-Propositional Model (2006). On this view, there is a substantive theoretical and evidential basis for the claim that 'Propositional reasoning … is assumed to be capable of altering associative representations under specific conditions' (Deutsch and Strack 2010: 65), and propositional representations and assessments of truth function within a reflective system, (System Two), taking some working 'data' from associative memory. (Recall also, Mandelbaum's earlier argument, that propositional structures *do inhabit* the System One domain and System Two processes are present and active within the supposedly exclusive world of System One). Such *interaction* is surely suggestive of reasons-responsiveness within the semicompatibilist model and will be considered in Part III. Further, such interaction between System One and System Two processes suggests how changes in implicit measures *can* translate into changes in explicit measures and behaviour. So, importantly, Strack and Deutsch present a model that *enables* understanding of the relationship between changes in implicit measures and changes in explicit measures and/or behaviour.

I will describe the mechanisms that can mediate the influence of implicit social cognition within the Strack and Deutsch interactive model. The *Handbook of Implicit Social Cognition* (Gawronski and Payne 2010), Chapter Four, *Building Blocks of Social Behaviour* (Deutsch and Strack 2010), specifically addresses the impact of implicit social cognition on behaviour (2010: 62). To really understand the nature of possible control of implicit bias related behaviour it is necessary to understand and connect the elements or mechanisms that *together* mediate the influence of implicit social cognition on behaviour. Such elements or mechanisms are represented in Fig 5.1 reproduced from *Building Blocks of Social Behaviour* (Deutsch and Strack 2010: 66). In harmony with Buckwalter's rejection of the first premise of the ability argument, it is possible to understand Fig 5.1 as essentially a model of control of behaviour, intuitively very plausible and supported by credible theoretical and practical research. This model integrates impulsive and reflective determinants that lead towards greater or lesser deliberative or reflective control of behaviour. Deutsch and Strack, at the outset, note the whole ' … paradigm of implicit social cognition rests on the notion that attitudes, prejudice, stereotypes, and the self may have an impact on behaviour that sometimes opposes beliefs and intentions' (2010: 62).

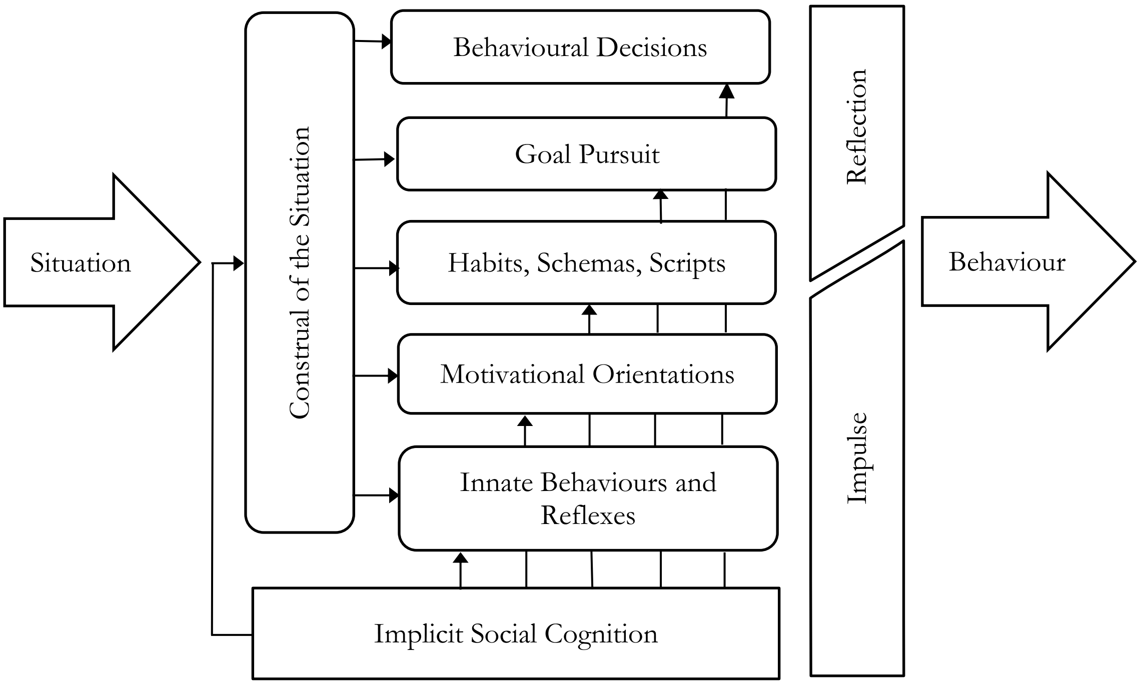
Situation
Construal of the Situation
Behavioural Decisions
Goal Pursuit
Habits, Schemas, Scripts
Motivational Orientations
Innate Behaviours and Reflexes
Implicit Social Cognition
Reflection
Impulse
Behaviour

Fig 5.1 Mechanisms that may mediate the influence of implicit social cognition (attitudes, stereotypes, self) on behaviour.

The main question addressed by Deutsch and Strack is, what are the *mechanisms* that facilitate, that mediate, the expression of such attitudes, prejudices and so on, in behaviour? Here, implicit social cognition is linked to basic ideas and assumptions already mentioned at some length (see Chapter 4); the idea of cognitive processes being subdivided into at least two groups, (implicit and explicit), that implicit social cognition research is generally based on indirect measures such as the IAT and finally, indirect and direct measures not only differ in procedure, but address different processes or systems. Further, as mentioned, the nature of implicit processes captured by indirect measures are usually considered associative in terms of underlying representations and activated automatically.[147] Recall, as discussed in Chapter 4, this view is not universally held; Deutsch and Strack (2010: 64) agree there are ambiguities concerning automaticity, such as reports of research participants able to control their immediate stereotypical responses.[148] Such ambiguities support the idea of an integrated reflective – impulse model, whereby reflective elements receive data from associative mechanisms of mediation, and also influence associative mechanisms.

A central requirement of semicompatibilism is moderate reasons-responsiveness so it is important to look further at the role of reason and reflection within this model of implicit bias / implicit social cognition. Further to discussion in Chapter 4, particularly *Attitude, Inference, Association: On the Propositional Structure of Implicit Bias* (Mandelbaum 2016), I will make some points concerning association and propositional representations and then mention some of the mechanisms that mediate the influence of implicit social cognition shown in Fig 5.1. As noted, on one hand, association does not have any truth value, no true state of affairs is inherent within the association itself, no factual link.

[147] If behavior issuing from implicit bias *is substantially* controllable then the moral distinction between behaviour originating from implicit attitudes and explicit attitudes appears to be failing. Clearly, if such behavior is substantially controllable then the first premise of the ability argument is undermined. This is contrary to considerable and credible counter opinion. For example, The Kirwan Institute for the Study of Race and Ethnicity is very clear concerning key characteristics of implicit bias (Staats 2013):

- Unconscious and *automatic*: They are *activated without an individuals' intention or control.*
- Pervasive: Everyone possesses them, even those avowing commitments to impartiality.
- Do not always align with explicit beliefs: Implicit and explicit biases are generally regarded as related but distinct mental constructs.

[148] For detailed discussion of automaticity see *Automaticity: A Theoretical and Conceptual Analysis* (Moors and De Houwer 2006).

Rather, association is formed and sustained by 'temporal contiguity and frequency of pairing' (Deutsch and Strack 2010: 65). Propositional activity takes place within the reflective system and subject to direct measures. On the other hand, Mandelbaum argues that implicit attitudes have more structure than (mere) associations; implicit attitudes are mental representations *with* propositional structure that function as unconscious beliefs (2016: 18). There is evidence supporting both positions and it is impossible within the limits of this Thesis to resolve this matter. That said, *interaction* under certain conditions between propositional and associative processes, rather than isolated operation, integrating impulsive and reflective determinants, *is* intuitively very plausible and importantly grounded in credible theoretical and practical research (Strack and Deutsch 2004; Gawronski and Bodenhausen 2006; Deutsch and Strack 2010: 65). Here, associative representations function as inputs to propositional processes and, (not shown in Fig 5.1), propositional reasoning can alter associative representations under certain conditions. This is an integrated approach in the sense that reflective and impulsive determinants work together, even though conflict may be present. See Appendix A, where these important issues are explored further. Next, I will look at some of the elements within the integrated model, mechanisms that may mediate the influence of implicit social cognition; motivation, opportunity to engage in reflective behaviour and construal of situations, (see Fig 5.1).

When lacking motivation and/or opportunity, the processes usually associated with indirect measures are particularly influential in determining behaviour, for example, habits, Schemas and Heuristics. Generally, when motivation and opportunity are present, processes associated with explicit measures have greater authority over behaviour, but the influencing bias of stereotypes and associations still exerts *some* pressure on an agent to act perhaps contrary to more reflective intentions. Influenced by implicit cognition, a possibly biased construal of a situation is available to various behavioural mediators together with *direct* connection to the influences of implicit cognition (see Fig 5.1). Implicit cognition thus has an indirect (via construal) and direct influence on behavioural mediators. One of the features of implicit bias that is often mentioned is dissonance between explicit views and beliefs, typically a sense of having liberal, egalitarian attitudes, and behaviour that issues from implicit attitudes such as a tendency to employ on the basis of gender rather than suitability. Deutsch and Strack

(2010: 70) call behaviour that issues from implicit attitudes irrational[149] and essentially associate it with impulsive System One processes. In terms of control, if impulsive determinants of behaviour have *completely* the 'upper hand' in certain situations, bypassing controlling/reflective/rational mechanisms, then there would be absence of control. While the integrative model of Deutsch and Strack (2010: 66) includes the possibility of reflective behaviour, there can be circumstances such as stress or tiredness (so lack of motivation) where this is almost impossible. However, tentatively, if implicit attitudes do have inherently more structure than (mere) associations, then responsiveness to reason and reflection may be more resilient to difficult (stress or tiredness) circumstances, *within* the Deutsch and Strack model.

The integrative approach concerning reflective and impulsive mechanisms evolves from and generally runs in harmony with Dual Process/System models described earlier (Chapter 4). However, importantly, for Deutsch and Strack reflective and impulsive systems 'operate interactively, serve different functions and have different conditions for optimal functioning' (2010: 72); impulsive systems (IS) are considered to drive implicit social cognition, whereas reflective systems (RS) complement IS, providing propositional representations based on what is active within IS, and importantly, have (unless the agent is under high stress), an *executive* function generating judgements and decisions that result in controlled behaviour.

To summarise, *change* in the sense of mitigating or perhaps eliminating implicit bias and *control* in the sense of exercising direct influence over biases and issuing behaviour, are different concepts. Buckwalter's references to changing implicit bias when control is the real issue, although I believe to be incorrect, has directed attention to the work of Strack and Deutsch, who propose an interactive unified System One and System Two model, important in understanding implicit attitudes and the mechanisms that mediate the influence of implicit social cognition (attitudes, stereotypes) on behaviour. Importantly, a model that offers the possibility of *control* of behaviour. This plausible and credible interactive model effectively closes down simplistic claims that S does not have the ability to *control* action Ø caused by implicit bias, because this model

[149] I believe 'irrational' is the correct description, rather than 'nonrational', as it captures that aspect of behavior that *deviates* from action that would be chosen rationally.

has as *part of* its structure mediating and reflective elements that have a controlling function on behaviour issuing from implicit attitudes. On the interactive view, *within* the mechanisms that determine behaviour are those providing rational executive control. While being 'within' and being *active*-within a mechanism are not necessarily the same thing, I believe the unified interactive model, incorporating reflective and mediating elements plausibly shifts the burden of proof *very* substantially towards those who claim an agent does not have the ability to control action Ø caused by implicit bias and so is not responsible.

Having covered much important ground since leaving Buckwalter and the ability argument, I will now continue, with discussion of the second premise; implicit attitude p *causes* action Ø. Buckwalter claims that current empirical evidence for a *causal* link between implicit attitude and behaviour is mixed; some studies show a modest relationship or no relationship, others claim implicit attitudes *do* make a unique contribution to *predicting* behaviour (2018: 8). But predicting behaviour and showing a causal link are clearly different matters; from Buckwalter there does not appear to be a *definitive* conclusion on the *causal* nature of implicit bias, other than the claim that there are currently doubts that implicit attitudes are a significant cause of behaviour, and so the second premise is not adequately supported, but future research may completely change this situation (2018: 11). To consider this question further and reach a clearer understanding of implicit bias in terms of its possible causal relationship with behaviour as reflected by premise 2 of the ability argument, (implicit attitude p causes action Ø), it is necessary again to look deeper into the nature of implicit attitudes and behaviour.

To look with particular interest at causation, I refer to two texts that together offer a comprehensive and very credible exploration of this aspect of implicit cognition: *Then a Miracle Occurs: Focusing on Behaviour in Social Psychological Theory and Research* (Agnew *et al.* 2010) and *The Implicit Mind: Cognitive Architecture, the Self, and Ethics* (Brownstein 2018).

From *Then a Miracle Occurs: Focusing on Behaviour in Social Psychological Theory and Research*, Chapter 6, Unconscious Behavioral Guidance Systems (2010: 1-36), Bargh and Morsella argue that within the unconscious, causation is active and 'there are a multitude of behavioral impulses generated at any given time from our unconsciously operating motives' (2010: 17). Description of the unconscious as a behaviour guidance system and *source* of action is plausible, for example, in an evolutionary context where the unconscious acts as source and guidance system for fast approach or avoidance. The essential and unequivocal claim expressed by Bargh and Morsella is that 'Social cognition research over the past quarter century has confirmed … unconscious processes for adaptively guiding human behavior existed prior to the advent of consciousness and continue to *generate* behavioral tendencies today' (2010: 16). A further point is made, that ' … evaluation and attitude activation, social perception, and goal pursuit (have) been found to be directly connected to behavioral tendencies, without any *need* for conscious intention or awareness in the production of these adaptive behaviors' (added emphasis 2010: 16). The claim that there are causal links between largely unconscious processes and behaviour is plainly made, but with one necessary clarification. Bargh and Morsella do not take *evaluation* out of the process that leads to behaviour, rather situate this activity in the unconscious rather than the conscious reflective domain shown in Fig 5.1. While it is difficult to understand how this could be possible, nevertheless, in terms of *causation* there is in Bargh and Morsella's opinion, no doubt that there *is* a link between unconscious processes and behavioural tendencies.

The Implicit Mind: Cognitive Architecture, the Self, and Ethics (Brownstein 2018) is described by Neil Levy '… as not the last word on the topic, but … the state of the art today' (2018). In light of Brownstein's reputation in the field of implicit cognition, for example, as joint editor and contributor to *Implicit Bias and Philosophy, Volume 1 and 2* (Brownstein and Saul 2016a and 2016b) and author of *Implicit Bias* (2016b), the claims and arguments relating to causation and implicit bias expressed within *The Implicit Mind: Cognitive Architecture, the Self, and Ethics* are clearly important and relevant within this discussion. This work is substantial, complex and subtle, developing a description of implicit attitudes 'that deserve to be considered a unified and distinct kind of mental state' (Brownstein 2018: 98), distinct from reflexes, mere associations, beliefs, dispositions and aliefs.[150] In terms of causation, I believe there is no doubt that Brownstein considers implicit attitudes have causal properties. Towards the end of

[150] For a detailed description of 'aliefs' see *Implicit Attitudes and the Architecture of the Mind* (Brownstein 2018: 85).

Chapter 3, Brownstein describes negative and positive *outcomes* of implicit attitudes whereby implicit attitudes can 'go awry from the perspective of moral or rational normativity' as in the case of implicit bias that manifests in irrational and immoral ends. However, sometimes implicit attitudes get it 'right' by *providing action guidance* in ways that our reflective beliefs and values do not, such as spontaneous reactions and decisions in athletic and artistic domains (following Brownstein 2018: 97). The notion of guidance is integral to, perhaps at the heart of, this model of implicit attitude. For example, when describing the difference between aliefs and implicit attitudes Brownstein says '… implicit attitudes have a pro tanto[151] guiding function … *signalling* to the agent to continue to attend or behave in some particular way' (Brownstein 2018: 94). Is the meaning of 'guiding' in this context the same as 'causing'? The term 'causes' within Premise 2, (Implicit attitude p causes action Ø), concedes no other necessary factors; implicit attitudes are sufficient for action Ø to happen and it will happen. Guidance, help and advice can be ignored or accepted, along with other factors that (may) result in action. The conclusion to be drawn from Bargh and Morsella and Brownstein is that there is a causal connection between implicit attitude p and action Ø, however, the nature of that connection should not be expressed simply as implicit attitude p causes action Ø. The previously mentioned model shown as figure 5.1 is an example of the complexity of interaction between implicit attitude and behaviour; that there are causal links between largely unconscious processes and behaviour seems irrefutable, but the nature of such causal links is an ongoing subject of empirical work and scholarship. Certainly, the nature of the complex link between implicit attitudes and behaviour is such that Premise 2 insufficiently describes the mechanism whereby implicit attitudes exercise their influence on behaviour. However, it is correct to say, I believe, that based on Bargh and Morsella and Brownstein, implicit attitudes *are* in certain circumstances a vital element in the causal process that leads to behaviour. Described in Appendix B (page 258), an alternative position is expressed by O'Connor; implicit bias is a contributor to a collection of *motivators*, but not in a direct causal role. Such motivational states could

[151] If a reason favours my doing something, then I have a 'pro-tanto' reason to do it; it is pro tanto *to that extent* right for me to do it.

include sufficient reasons for acting 'of which I am entirely *unconscious*' (O'Connor 2013: 235)).

The third premise from the ability argument, (if S is morally responsible for Ø, then S has the ability to control Ø), continues to be the subject of considerable debate and scholarship, for example, recall the important work of Frankfurt from Chapter 2, where agent responsibility is plausibly present *without* control in the broader regulative sense of making a choice between genuine alternative possibilities. This premise also recalls the plausible and well-known claim that ought implies can. For an agent to be morally responsible for an action, they must be *able* to fulfil (or to decline to fulfil) that action. In terms of the third premise, S is responsible for Ø implies S has sufficient control of the relevant circumstances relating to Ø to fulfil Ø. Buckwalter's critique of the third premise of the ability argument, that if S is morally responsible for Ø, then S can control Ø, is in the form of examples, where there is absence of ability but plausibly there is responsibility. The cases involve the familiar situation of an agent deliberately taking an action that knowingly will remove their ability to perform a particular task in the future. For example, deliberately leaving home late to avoid a meeting, so at the time the meeting should start it is beyond the ability of the agent to attend. At the later time, the agent through their own deliberate action does not have direct control or direct responsibility, but clearly because their earlier action was an expression of *indirect* control the agent now has indirect responsibility for absence. Such examples claim to show, contra the third premise, the possibility of moral responsibility for actions that are beyond an agent's direct control, by virtue of being indirectly responsible. Buckwalter claims it is likely that agents *are* responsible for their implicit bias related actions in the sense that they are at least indirectly responsible, for example, by failing to be sufficiently aware of implicit bias issues, even though such information is readily available.

The claim that failing to be sufficiently aware of implicit bias issues when such information is readily available (possible culpable ignorance) can lead to *in*direct responsibility for behaviour appears very plausible, but to what extent *are* we morally responsible for investigating (and eradicating) our biases? I will describe an approach to this question, comment on the unlikely case of an agent who is completely unaware of implicit bias and therefore takes no prior action, look at the transfer of responsibility principle and then conclude discussion of the third premise (page 132).

Recall Erin Beeghly, *Bias and Knowledge Two Metaphors* (2020: 77-98). This Paper begins with the question, 'if you care about securing knowledge, what is wrong with being biased?' The answer is developed using, (as the title suggests), two metaphors, bias as fog and bias as shortcut. It is impossible to adequately summarise Erin Beeghly's far-reaching and essential Paper. Erin Beeghly argues clearly and persuasively that we are deeply morally responsible for investigating (and eradicating) our biases from the perspective of justice, but what particularly resonates in this Paper is the need to 'look out' into the world at the social dimensions of bias in terms of privilege and oppression that do not receive appropriate attention when internal cognitive processes, motivation and cognitive overload are the main focus.

Lindsey and Bradley Rettler in their significant Paper *Epistemic Duty and Implicit Bias* (2019) argue that people have an epistemic duty to eradicate at least some of their implicit biases. Moral duties and epistemic duties are clearly different, but Lindsey and Bradley Rettler look carefully at the close parallels between them. Implicit biases are epistemically harmful because they interfere with, or block, other epistemic duties such as having opinions and views that are true, based on clear and good reasoning. Given that implicit biases affect our ability to achieve other epistemic duties they should be eradicated. There is a clear parallel with moral duty;

1. We ought to fulfil our moral duties.
2. Having implicit biases prevents us from fulfilling our moral duties, so
3. We ought not harbour implicit biases.
4. If we can eradicate our implicit biases, then we (morally) ought to.

The moral and epistemic arguments have the same structure. Both rely on the connection between implicit biases and other things - moral or epistemic duties (2019: 8). Looking at Lindsey and Bradley Rettler's argument in a little more detail, it is argued that agents have an *epistemic duty* to eradicate implicit biases that have negative epistemic impact; for example, when implicit biases get in the way of knowing the truth or having correct beliefs about something. Agents have such a duty if they have the right kind of *control* over implicit biases and blameworthy if having such control, they do not eradicate them. The right kind of control is available and referred to as indirect reflective control.[152] The argument runs like this: Implicit biases can have a negative epistemic impact, for example, when they cause formation of false and unjustified beliefs, they function as knowledge blockers or by preventing other epistemic duties such as believing what is true and rejecting what is false (Rettler and Rettler 2019: 8). Implicit biases lead to bad outcomes but are also epistemically bad in themselves; they may be simply incorrect (assuming implicit biases are truth apt) or call up false associations. What is meant by 'the right kind of *control* over implicit biases'? Without the right kind of control there is no duty to eradicate implicit bias. The popular view of implicit bias as 'below the radar of consciousness' suggests as discussed, that control is impossible; Rettler and Rettler, while accepting the claim that direct control of implicit bias *is* impossible, maintain that *indirect* control is available and sufficient, granting the right kind of responsibility to an agent who is subject to legitimate blame when failing in the duty to eradicate implicit bias. Indirect control of *beliefs* is described as having 'control over whether she believes that *p* iff she can actively engage in critical reflection that causally influences whether or not she holds the belief that *p*' (Rettler and Rettler 2019: 13); reflecting on the reasons for believing *p* and changing our beliefs as necessary, considering such reflection. For *implicit biases*, Rettler and Rettler argue that the necessary indirect control is 'even more indirect than our control over belief' (2019: 14) in the sense that eradicating implicit bias is more successful when, for example, reflecting on positive *arguments* that support efforts to recruit more women into academic philosophy rather than reflecting on reasons not to have implicit bias concerning women philosophers. Alternatively, Holroyd (2012: 288) argues manifestation of implicit bias is influenced by *explicit* beliefs. Reflecting on explicit beliefs and values has an indirect influence on implicit values and biases. There are other indirect ways to control implicit bias described shortly.

[152] There are two ways in which we have indirect reflective control over whether we harbour implicit biases. One is that we can actively engage in reflection on particular arguments, as well as various beliefs and values that we hold, and this reflection causally influences whether we harbour various implicit biases. The other is that we can actively engage in reflection on various techniques that help eradicate implicit biases, and this reflection makes a difference to whether we engage in these activities, which in turn makes a difference to whether we harbour various implicit biases. Again, it's not that agents have to actually reflect in these ways to be said to have indirect reflective control; but rather it's that agents are capable of such active reflection (Rettler and Rettler 2019: 17).

So, agents have an *epistemic duty* to eradicate implicit biases that have negative epistemic impact if they have the right kind of control; to the extent that indirect reflective control is lacking, an agent is not blameworthy for failing in their duty to eradicate implicit bias. There is a *moral duty* to eradicate implicit biases that are blocking our ability to fulfil other moral duties. If the right kind of control is possible then, as in the case of epistemic duty, we ought to exercise this control or be subject to blame.

What if an agent is completely unaware of implicit bias issues and therefore takes no prior action (or inaction) that would make later responsibility plausible? In such circumstances surely there cannot be responsibility for their implicit bias related actions? Buckwalter highlights the difference between assessing moral responsibility and assessing blame. It is suggested that an agent acting in complete ignorance may be responsible for implicitly caused behaviour while leaving open the question of blame given such excusing circumstances. The third premise is revised; for S to be morally *blameworthy* for Ø, then S must have the ability to control Ø. This appears at first look to be a difference without great substance, the notion of blameworthiness now carries all the moral weight. Buckwalter points out, in terms of implicit bias related behaviour, accepting there is responsibility while keeping open discussion of the more emotive allocation of blame may be helpful, for example, in public implicit bias mitigation strategies (2018: 17). That said, Buckwalter's distinction between responsibility and blameworthiness does not illuminate anything *significant*, such as interesting new insights or directions of argument. Madva presents a nuanced position on this issue in Section 6 Conclusion of *Implicit Bias, Moods, and Moral Responsibility* (2018):

> While I agree that we should not necessarily saddle individuals with '-ist' labels that portray them as horrible people for possessing and expressing implicit biases, it is a mistake to conflate sanctimonious name-calling with the view that implicit discrimination is often worthy of blame, broadly construed. Blame is not so blunt an instrument. We can acknowledge the failings of others and ourselves to live up to our commitments without calling the sincerity of those commitments into question. In many cases, we can insist that individuals bear a legitimate degree of responsibility and blame, even if they lack perfect awareness of what they do. If it is ever strategically unwise to lay blame, then the upshot is not to jettison implicit bias from the sphere of moral responsibility; the upshot is to take great care in locating it properly within that sphere. (2018: 71)

If lack of ability to perform an action or to refrain from an action mitigates blame then what is the meaning, role or definition of responsibility? An alert, competent and sober driver will *feel* extremely responsible for harming a pedestrian who suddenly steps into the road even though it was not their fault, and they were not blameworthy. This issue is considered later during discussion of moral luck where it is suggested that the driver experiences a feeling of profound *regret* rather than responsibility. It is unlikely that separation of often emotive and controversial concepts of responsibility and blame could be sustained, particularly when discussing or implementing implicit bias mitigation strategies, a topic inherently emotive in the sense that participants may already be feeling hostile to the suggestion that they have implicit bias that requires mitigation.[153]

Concluding discussion of the third premise, if S is morally responsible for Ø, then S has the ability to control Ø. This is essentially a sound claim if 'control' is understood as guidance control, so actual alternative pathways need not be available, the truth of determinism left as an open issue and moral responsibility retained. Buckwalter's critique of the third premise in terms of direct and indirect responsibility is noted but I do not believe it has a significant impact, for example, on John Martin Fischer's discussion of tracing as an integral and important part of the historical notion of guidance control and responsibility (1998: 195), paying careful attention to cases where responsibility is intuitively very questionable, such as compulsive behaviour (1998: 48). As mentioned, these themes are present in the development of semicompatibilist control and responsibility, discussed in Chapter 3.

As mentioned at the beginning of this section (page 128), to complete discussion of Buckwalter and the ability argument I will look at the transfer of responsibility principle, or control transfer principle (CTP):

CTP: If p causes Ø then the inability to control p entails the inability to control Ø.

Buckwalter interestingly reformulates the CTP in terms of behaviour and implicit attitudes, calling the reformulation the implicit control transfer principle (ICTP):

[153] See also *Responsibility for Implicit Bias* (Holroyd 2012: 298) for detailed discussion of blaming and holding responsible.

ICTP: If implicit attitude p causes Ø then the inability to control p entails the inability to control Ø.

Buckwalter rejects the ICTP; the transfer of lack of control of implicit attitudes to lack of control of behaviour, for three reasons. First, the claim that inability to control p entails the inability to control Ø when Ø is caused by p is *not* universally accepted. Second, more importantly, there is insufficient supporting evidence for the claim that implicit attitudes *are* uncontrollable and ' … it is unlikely … that implicitly biased agents lack so-called "responsibility-level control" of their behaviour' (2018: 20). (It is also suggested that if implicit attitude p causes Ø, implicit attitudes make only a small contribution to behaviour and so are not important).[154] Finally, rejection of ICTP is based on the claim that 'controllable behaviour *can* follow from uncontrollable states' (2018: 22).

Summarising, the ability argument maintains implicit attitudes cannot be controlled and are a source of discriminatory behaviour that, based on the implicit control transfer principle, are not subject to agent control, or responsibility.

1. S does not have the ability to control implicit attitude p.
2. Implicit attitude p causes action Ø, therefore, S does not have the ability to control action Ø (by the Transfer of Responsibility Principle).
3. If S is morally responsible for Ø, then S has the ability to control Ø.
4. Therefore, S cannot be morally responsible for Ø.

Conclusions so far, taking each premise in turn;

1. S does not have the ability to control implicit attitude p. There is insufficient supporting evidence for the claim that implicit attitudes cannot be controlled. From Deutsch and Strack (2010), an interactive model incorporating reflective and mediating elements provides rational executive control, and as a minimum, plausibly shifts the

[154] An interesting point is suggested here - the idea of a causal link that makes only a small contribution to behaviour. How is this to be understood? Surely, even if the 'amount' of behaviour issuing from an implicit attitude is relatively small, this does not thereby make such a causal implicit attitude any less important, given the sorts of areas where implicit attitudes and bias are often expressed.

burden of proof *very* substantially towards those who claim an agent does not have the ability to control action Ø caused by implicit bias and is so not responsible.

2. Implicit attitude p causes action Ø. There does not appear from Buckwalter to be a definitive conclusion on the *causal* nature of implicit bias. However, on further consideration based on Agnew *et al.* (2010) and Brownstein (2018) the complex link between implicit attitudes and behaviour is such that Premise 2 although essentially true, is an insufficient description of the mechanism whereby implicit attitudes exercise their influence on behaviour. However, I believe, based on Bargh and Morsella and Brownstein, it is correct to say that implicit attitudes *are* in certain circumstances a vital element in the causal process that leads to behaviour.

3. If S is morally responsible for Ø, then S has the ability to control Ø. Buckwalter presents cases where an agent acts to ensure at a later time control is impossible, yet still plausibly retains responsibility. The premise, if S is morally responsible for Ø then S has the ability to control Ø is revised; for S to be morally blameworthy for Ø, then S must have the ability to control Ø. The third premise is essentially sound. (It is understood that Frankfurt type examples could be presented to counter this claim). While Buckwalter's critique of the third premise in terms of direct and indirect responsibility is noted, I do not believe that it changes the way forward within the overall discussion of control and responsibility; these issues are discussed in greater depth in other areas of the Thesis.

The validity of the ability argument is challenged by disputing the Control Transfer Principle; if p causes Ø then the inability to control p entails the inability to control Ø. Various challenges are presented, perhaps most plausible is the claim that based on counter examples 'controllable behaviour *can* follow from uncontrollable states' (Buckwalter 2018: 22).

4. Therefore, S cannot be morally responsible for Ø. Buckwalter comprehensively challenges the truth of the premises and validity of the ability argument, concluding that 'Pending future evidence, the rejection of these premises undermines the ability argument against moral responsibility (for implicit attitudes and related behaviour) (my addition 2018: 23).

The title of this chapter is Implicit Bias and Control, therefore, to be clear, drawing on Buckwalter (2018) and Deutsch and Strack (2010), the first premise, that S does not have the ability to control implicit attitude p, is rejected. Further, a contrary position is posited, that an agent can control implicit attitudes *and* related behaviour. Central to this claim is the interactive model (Fig 5.1), a configuration that is plausible and supported by significant empirical and theoretical research. There is behavioural decision-making present within this model, the possibility of responsiveness to reasons, (essential to control and responsibility within the semicompatibilism of John Martin Fischer). While not explicit within the model outline, there is interaction present in both 'directions' in the sense that Reflective and Impulsive Systems influence and are influenced by each other (Deutsch and Strack 2010: 73). Such interaction suggests an integrated implicit bias mitigating strategy based on reason and argument *and* positive association could be particularly effective.

The Holroyd and Kelly Approach

I will now consider Holroyd and Kelly's (2016) alternative approach to implicit bias and control. Is the disposition to behave in ways influenced by implicit bias part of a person's character and so subject to moral evaluation? Responding to this question Holroyd and Kelly believe it is necessary to consider if an agent can *control,* (in some clearly defined way), such behavioural dispositions.[155] Clark's 'ecological control' is said to provide the necessary control, supporting the claim that character-based evaluation of such behavioural dispositions is justifiable. So, individuals have relevant control and responsibility for implicit bias, where implicit bias and the disposition to act in accordance with its influence are unified within character and can be legitimately morally appraised. In other words, implicit bias is an expression *of* our true self and something to be appropriately controlled *by* our true self. Clearly, this challenges popular and some expert opinion, i.e., that individuals do not have control of behaviour that takes place because of the influence of implicit bias. Holroyd and Kelly's approach will be described along with five forms of control. The claim that implicit bias reflects an agent's character and thereby can be legitimately morally appraised may be expressed as follows;

[155] Recall the distinction between direct and indirect control should be noted; direct control at the point of behaviour and indirect control, as suggested, for example, by Clark's ecological control, at a time prior to the point of behaviour.

1. *If* an agent has the relevant control and responsibility for implicit bias, *then* implicit bias reflects an agent's character.
2. Individuals *do* have relevant control/responsibility for implicit bias.
3. Therefore, implicit bias reflects an agent's character, (and can legitimately be morally appraised).

Crucially, what is the nature of the 'relevant control' over implicit bias that sanctions evaluation of an agent's character and what is the role and implications of luck in formation of character? Jules Holroyd and Daniel Kelly suggest there are five forms of control to consider; direct (see also page 128), unified agency and reflective, alienation and evaluative, intervention, and ecological (2016: 108). It will be seen that ecological control offers the necessary means whereby an agent *can* take ownership of implicit bias as part of their character. I will now look briefly at each of these forms of control.

Direct control

Direct and immediate control of bias by the agent on becoming aware of certain behaviour is impossible under the basic associationist model; breaking associations that have built up over perhaps many years takes time. By contrast, a rational-responsive concept of implicit bias suggests that direct and fast control should be possible,[156] but, as Eric Mandelbaum asks, '… if evidence can flip implicit attitudes, why don't aversive racists drop their implicitly biased attitudes?' (2016: 23). Mandelbaum offers an explanation, while retaining a rational-responsive concept of implicit bias, suggesting that when unconscious attitudes change it is clearly not a straightforward process; 'the more important the content is, the more beliefs and connections it will have. To overturn implicit biases, it will be necessary to tackle all of these (many) different representations …' (2016: 23). On this account, the notion of direct control does not seem attractive as a means of implicit bias related behaviour control.

[156] See also Levy, *Implicit Bias and Moral Responsibility: Probing the Data.* For example, 'The causal processes whereby implicit attitudes modulate behavior and decision-making are opaque to introspection, ensuring that we lack insight into what influence they have on our perceptions and judgments, and that there are no reliable means of modulating or inhibiting this influence' (2017: 5).

Unified agency and reflective control

In Part III (see page 197) Levy's classic Paper *Implicit Bias and Moral Responsibility: Probing the Data* (2017) is discussed in terms of the nature of implicit bias and moral luck. At this point I will refer to Levy's Paper in terms of its exposition of the central issues of responsibility and implicit bias, including detailed consideration of control and reasons-responsiveness. Levy concludes that implicit attitudes 'are not beliefs but patchy endorsements: endorsements because they have sufficient propositional structure to have truth conditions, but too patchy to be genuine beliefs' (2017: 8). Levy argues that such patchiness undermines patterned reasons-responsiveness, control and responsibility. That responsibility is so undermined seems to Levy to be counterintuitive, implausible, and so continues by describing an alternative approach referred to as the attributability view. On this view, 'agents are responsible for attitudes that properly belong to them, and for actions caused by such attitudes' (Levy 2017: 18). However, after much consideration Levy says:

> It seems that on the attributability account,[157] agents *can't* be morally responsible for actions caused by their implicit attitudes when those actions have a moral character that diverges from the character they would have had were their explicit attitudes controlling their behavior – for under those conditions implicit attitudes do not belong to the *real self.* (added emphasis 2017: 14)

In other words, implicit attitudes do not sufficiently *belong* to agents such that they are part of their real selves. It should be noted that the concept of 'real self' is problematic in the sense previously noted, that to a greater or lesser extent the real self is constituted by luck and the environment generally and so not under our full control. Therefore, behaviour that has the true self as source is not fully under agent control and so not fully responsible behaviour. Levy's overall account is *considerably* more nuanced than such brief description allows, but the conclusion is clear; 'if control, or attributability, or both, are necessary conditions of moral responsibility, agents are *not* directly responsible for

[157] Moral responsibility is usually framed in terms of control or attributability. An agent is morally responsible for an action, or for the consequences of an action, iff they exercise 'freedom-level' control, (deliberate control, exercised in the service of an explicit intention). Or, framed in terms of attributability, where an agent is morally responsible for an attitude or an action iff it is appropriately *attributable to* the agent (following Levy 2017: 4).

actions that have a moral character due to their implicit attitudes' (2017: 14). Levy's argument is based on the claim that the nature of *ex*plicit attitudes (generally) is such that they form a coherent whole that retains a reasonable degree of continuity over time, by contrast implicit associative attitudes are *not* overall coherent and *not* controllable in a way that allows unification of agency and expression of who the agent is. Hence, implicit associative attitudes are not the proper objects of moral assessment. Jules Holroyd and Daniel Kelly disagree for two reasons:

> First, we see no reason to suppose that implicit, associative processes in general, as a kind of mental structure, cannot help contributing to the unification of an agent. [...] Second, even if we accept Levy's claim that such states disrupt or fail to contribute to unified agency, this does not entail that such states are not candidates for moral evaluation in an agent who meets some threshold of unity by other means. (Holroyd and Kelly 2016: 113)

On Holroyd and Daniel Kelly's view, even *if* it is accepted that implicit associations are not *unified* within an agent, this does not entail implicit associations contribute nothing to who the agent is or cannot be to that extent subject to evaluation. Referring to Moskowitz and Peizhong's Paper *Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control* (2011), Holroyd and Kelly argue there is evidence that automatic *implicit* processes, *outside of the awareness* of reflective agency, do work towards bringing behaviour into line with explicit values; they act in a unifying way, not alien to the self or character. Moskowitz and Li conclude; 'It (stereotype control) is something we proactively[158] engage, outside of conscious awareness, to help produce desired cognition in the first place, even inhibiting unwanted thoughts before they occur' (2011: 115).

Moskowitz and Li's Paper presents fascinating conclusions relating to this *unifying* aspect of implicit cognition that merits full presentation; the essential claim is that 'An individual can control stereotyping without knowing a stereotype or a goal exists. Conscious awareness of goals is not required. One's wants, even implicit wants, can direct thoughts' (2011: 114). This sounds too good to be true, but Moskowitz and Li present substantial and compelling empirical and theoretical supporting evidence.

[158] The term 'proactive' as used here means 'Control ... exerted on stereotype activation at the first step of the process. This is a *proactive* strategy of control, one focused at the level of basic social - cognitive functioning ... ' (Moskowitz and Peizhong 2011: 104).

Essentially, it is claimed there will always be a problem trying to inhibit stereotypical responses based on a Dual System model of conscious processes. Attempts to stop *already activated* stereotypes from influencing behaviour contrary to an agent's explicit goals by System Two intervention are not effective. Simply expressed, by the time System Two processes intervene it can be, and often is, too late. Based on their substantial empirical work, Moskowitz and Li describe a more successful strategy, a *proactive* form of control that prevents a stereotype from ever being retrieved from memory, even though perceivers have categorized a person to a social group. Moskowitz and Li claim that 'Control can be exerted on stereotype activation at the first step of the process. This is a proactive strategy of control … arguing that (appropriately set) goals disrupt *the activation* of stereotypes' (2011: 104). It is claimed that goals are not always consciously selected and given that cognitive processes serve a goal, whether stereotype activation occurs is thus dependant on what goal the individual is implicitly pursuing. A nonconscious goal to be egalitarian directs selective attention to goal relevant stimuli in the environment that an individual is not consciously able to detect. Goals incompatible with stereotyping can be primed and held by the individual outside of their awareness.[159] (following very closely Moskowitz and Peizhong 2011: 104 - 105).

It should be mentioned again that, although not described here, these theoretical claims and discussion are based on and supported by empirical work detailed within Moskowitz and Peizhong's Paper. The fundamental conclusion is that automatic implicit processes are working in a unifying role, bringing behaviour into line with explicit values. This sense of unity whereby implicit bias, the disposition to act in accordance with its influence and the implicit countering tendency to bring behaviour into line with explicit values, lends weight to the claim of a unified character legitimately subject to moral appraisal.

[159] The obvious question arises, *how* can a person adopt a goal to inhibit a stereotype-based response, (without awareness or conscious intent to inhibit the stereotype at the time such inhibition occurs); how is it possible to acquire an implicit goal that is consistent with an explicit, (for example, egalitarian), World view? The experimental methods of Moskowitz & Peizhong describe how under lab conditions this is facilitated, but it is not clear how this would be achieved outside such controlled conditions.

Alienation and evaluative control

The term alienation is used to describe implicit bias and its claimed *unresponsive* nature to an agent's explicit evaluative attitude. The agent is said to be alienated from their implicit attitudes. From this idea, various positions emerge, for example, that implicit biases are *not* part of the real self and are exempt from moral assessment. Alternatively, from Glasgow's[160] account, implicit biases while not part of an agent's moral character due to alienation, or lack of evaluative control, nonetheless influence the moral evaluation of an agent and their actions (Brownstein and Saul 2016b: 37-61). Holroyd and Kelly (2016: 114) agree that alienation does not entail any reason to exempt an agent from evaluation of their alienated attitudes and biases but disagree that alienation entails that implicit biases are not part of who the agent is.[161]

Intervention control

If an agent cannot inhibit or intervene in implicit bias influenced actions due to simply being unaware of their presence, as noted previously, this naturally raises considerable doubts over granting control, attributability and responsibility to that agent. Intervention control occurs when an agent inhibits or redirects their own behaviour when it is sensed that it is deviating from explicitly held views and commitments. Holroyd and Kelly agree that agents often cannot assert this type of control over implicit bias related behaviour. However, such a limited concept of control, where an agent 'steps in' quickly to inhibit or redirect their own behaviour, bringing into line perhaps with explicitly held views and commitments, is surely a too exacting requirement. Ecological Control provides a significantly more nuanced approach offering a plausible, perhaps compelling, description of unifying control that supports the idea that implicit biases are part of an agent's character and so subject to evaluation.

[160] See Joshua Glasgow, Alienation and Responsibility, *Implicit Bias and Philosophy Volume 2: Moral Responsibility, Structural Injustice, and Ethics* (Brownstein and Saul 2016b: 37-61) for comprehensive treatment of this idea.

[161] Recall earlier comments concerning the concept of 'real self' as potentially problematic in the sense that it is arguably to a greater or lesser extent constituted by luck and environment generally and so not under our full control.

Ecological Control

Andy Clark's description of ecological control (2006) is subtle and quite radical in its approach. Indicative of Clark's general position is the following:

> There is no self, if by self we mean some central cognitive essence that makes me who and what I am. In its place there is just the 'soft self': a rough-and-tumble control-sharing coalition of processes – some neural, some bodily, some technological – and an ongoing drive to tell a story, to paint a picture in which 'I' am the central player. (2006: 23)

The notion of the mind having a fluid boundary, not coinciding with the biological boundary of the organism, appears at first glance to be from the realm of science fiction rather than philosophy, but Clark argues convincingly that particularly human beings incorporate, draw in and unify, features of the environment and themselves when acting and problem solving, describing human beings as 'ecological controllers' (2006: 5). Ecological control is a top-level control, not the micromanaging of finer operating details of a system. Holroyd and Kelly summarise Clark's position:

> … one central feature of human agency involves supplementing the internal sub-personal mechanisms that guide behaviour by engineering their world, calibrating 'external' sub-personal structures so that they help simplify cognition and bring out the kinds of behaviours and outcomes to which they aspire. (2016: 118)

How does this idea of ecological control connect with mitigation of implicit bias? While agents probably lack direct intervention control over implicit biases, it is plausible that *some* degree of malleability of implicit bias attitudes is possible (Devine 1989), (Blair 2002), (Moskowitz and Peizhong 2011). If it is accepted that change is possible, then Clark's model of ecological control can be helpful in framing some of the change motivating options available to the agent. For example, using environmental props to guide cognitive processes such that implicit biases are weakened. This could take the form of counter conditioning, such as thoughtful exposure of agents to positive images of admired black celebrities around the work-place, where the work-place on Clark's account is an extended cognitive field (2006: 25). The essential point from Holroyd and Kelly, based on Clark's account, is that features of the environment can be managed to influence cognitive processes in predetermined ways, including automatic features of our mental processes. *If* such influencing effects of the environment are potentially available for positive manipulation, it seems a very plausible claim that an agent's implicit biases are at least to this extent under their control, with attendant responsibility. (On page 118, it was noted that changeable and controllable are not simply interchangeable terms. Here, the subtle and detailed approach by Andy Clark counters any notion of simplicity and obvious concerns of change of implicit attitudes as a form of control). The agent must be aware of implicit bias and have some knowledge of ecological control, control that in many cases can take the form of quite straight forward strategies. Holroyd and Kelly describe ecological control as 'the structuring of one's environment and cognitive habits such that autonomous processes and sub-systems can effectively fulfil one's person-level goals' (2016: 130). It is claimed that ecological control offers the possibility of character-based evaluation of the agent's *implicit* attitudes because in this important ecological sense the agent *can* have control over such mental states. Not by direct intervention into the automatic process as conventionally understood, but with a broader notion of control by purposefully manipulating an agent's environment.[162] [163]

Summarising the above greatly condensed description of Holroyd and Kelly's[164] Paper: (i) There is an important sense in which individuals have control over implicit biases. Such control is commonplace in our exercise of agency. (ii) Ecological control is the structuring of one's environment and cognitive habits such that autonomous

[162] In their short, intense and very interesting Paper, Phia Salter, Glenn Adams and Michael Perez *Racism in the Structure of Everyday Worlds: A Cultural-Psychological Perspective* (Salter, Adams, and Perez 2018), argue that to understand racism a cultural-psychology framework is necessary. This Paper is difficult to summarise, but an essential point is the importance of culture in a wide variety of aspects that reinforces racists attitudes that in turn facilitate a racist culture. Relating to Ecological Control, although not mentioned explicitly within the text, the authors conclude that 'rather than attempt to control expression of culturally constituted, *individual* bias, a more effectual use of personal agency may be to reconstruct worlds that promote antiracist tendencies' (added emphasis 2018: 153). See also Ayala-López and Beeghly, Explaining Injustice: Structural Analysis, Bias, and Individuals, *An Introduction to Implicit Bias* (2020: 211-232).

[163] Such indirect intervention, control, and so responsibility (or any intervention) surely depends on at least some meaningful awareness of implicit bias. A response to this objection may be found in *Responsibility for Implicit Bias* (Holroyd 2012: 292). Essentially, it is argued that 'it is not a necessary condition for responsibility that individuals are aware of the influence of certain cognitive states on their decisions and actions' (2012: 294).

[164] With minor changes, my summary is based on Holroyd and Kelly's Paper, Section 5.6, Concluding Remarks (2016: 130).

processes and sub-systems can effectively fulfil one's person-level goals. There are two central implications of this claim: If agents have this kind of control, (and so responsibility, then subject to other necessary conditions being met, discussed below), implicit attitudes can be an appropriate target of character-based evaluation. If agents exercise this form of control, (and so responsibility, then subject to other necessary conditions obtaining), implicit attitudes can be properly regarded as part of 'who the agent is' - part of her character, which is as a whole subject to moral evaluation (2016: 130).

There is an important point to note. The argument that an individual *can* take and exercise ecological control over implicit biases is vulnerable to the following objection.

> … whether, for any individual, she can in fact exercise ecological control depends on whether she is aware of these possibilities (and indeed, aware of the phenomena of implicit bias, and that she may be affected by it). So the mere possibility of having ecological control is not sufficient for implicit biases to be considered as 'part of the agent' and hence morally evaluable. In addition to the control conditions, *epistemic conditions* must also be met as well. (added emphasis Holroyd and Kelly 2016: 127)

Holroyd and Kelly *accept* that ecological control can permit moral evaluation only if other conditions obtain, however it is claimed such an objection can be moved to the side. The vital claim and supporting argument are that implicit biases are part of who the agent is, and agents can be evaluated for being influenced by them because the agent has control over such mental entities. Holroyd and Kelly identify an important sense in which agents *can* have control (ecological control) over such mental entities (following closely Holroyd and Kelly 2016: 175).

Following some brief comments on the notion of 'who the agent is' I will bring together and summarise the ideas relating to implicit bias so far discussed, together with some additional points, in preparation for Part III.

The idea of 'who the agent is' or the notions of 'real self' and 'deep self' offer a perspective from which important issues such as agency, responsibility, free will and autonomy are obviously connected. Claims of responsibility, with associated praise or blame, are usually based on some notion of autonomy. It is intuitively felt that to act responsibly it is necessary to act autonomously; clearly without coercion, but also in some sense as an uncaused cause, whereby actions originate from 'within' the self and without any overriding 'external' influences. The notion of the self as an 'uncaused cause' is clearly problematic, not least because without a cause or influence at some stage it seems that actions occur randomly and so preclude responsibility. Typically, the cause, or agent causation, is understood as an emergent human (and some nonhuman animal) property, that while still ontologically physical, has the capacity to initiate causal chains in a top-down direction to lower levels of the body's hierarchical structure and so into the world. If responsibility does not entail autonomy, then how is responsibility and free will to be understood, if at all? Alternatively, if responsibility is to be meaningful then how is autonomy to be understood? From earlier discussions,[165] this is familiar territory, but as Susan Wolf notes,[166] within the free will debate focus is usually on determinism of some form rather than the connection between free will and personal autonomy (1993: 24). The essential objective noted at the beginning of *Freedom Within Reason* (Wolf 1993) is to understand and describe a sense of autonomy that would give necessary freedom of will and action to agents such that responsibility, with associated praise and blame, could legitimately be assigned. The problem is to identify and convincingly describe a self, (perhaps in terms of deeply held values), that, as suggested, is in some sense 'uncaused' (autonomous) and yet be a source *of* sequences of causation.[167] How is it possible or meaningful to say that a real or deep self exists exempt from external causal influences from the environment, from genetic determination or simply from luck? To be influenced by these things, in ways that shape behaviour is, it seems, potentially to relinquish autonomy in ways that could be relevant when assigning responsibility. At this point, the ontology of such a self remains obscure and mysterious.[168] For further

[165] See particularly David Hume page 19, Determinism page 37 and Wolf (1993: 27) page 59.

[166] *Freedom Within Reason* (Wolf 1993), particularly Chapter Two, presents a superbly clear development and critique of the real self-view.

[167] For detailed expression of these issues see Galen Strawson *The Impossibility of Moral Responsibility* (G. Strawson 1994), Clarke *On an Argument for the Impossibility of Moral Responsibility* (2005) and Hartman *Constitutive Moral Luck and Strawson's Argument for the Impossibility of Moral Responsibility* (2018), also available from <www.robertjhartman.com>.

[168] The hugely influential philosopher and historian Michel Foucault rejected fixed essence as personal identity. Foucault argued that the self is defined by a *continuing* discourse in a shifting communication of oneself to others. The classical view of identity is something that is inherent and fixed in some way or

discussion of this key issue see Appendix B Agent Causation, where agent causation and other models of agency are described in greater depth.

5.2 Summary

The aim of this chapter was to formulate a description of implicit bias to take forward into Part III. Recall that Part III considers whether behavioural expression of implicit bias is such that it is (or is not) subject to guidance control and hence whether the agent acted freely and responsibly (as those terms have been developed within the semicompatibilist model). If examination shows that issuing behaviour of implicit bias is *not* subject to guidance control, then the agent does not act freely and responsibly as articulated by John Martin Fischer's semicompatibilism. Such a conclusion would be contra to what has now been developed within Part II where compelling arguments show agents *are* responsible for behaviour that issues from implicit bias. If such a conflict occurs this is suggestive of a problem with the semicompatibilist model, and it may be possible to propose amendments or revisions to John Martin Fischer's semicompatibilism to take account of the issues brought to light by implicit bias.

Two important approaches to implicit bias and responsibility have been considered with reference to; *Implicit Attitudes and the Ability Argument* (Buckwalter 2018), *The Handbook of Implicit Social Cognition* (Gawronski and Payne 2010), Chapter Four, Building Blocks of Social Behaviour (Deutsch and Strack 2010) and *Implicit Bias, Character, and Control* (Holroyd and Kelly 2016). From examination of Buckwalter's critique of the premises of the ability argument I worked towards description of a plausible interactive Dual System model based on Deutsch and Strack (2010), a configuration that includes decision making and mechanisms that mediate the influence of implicit social cognition on behaviour. From Holroyd and Kelly I described a model of individual control over implicit bias and influenced behaviour: Ecological control requires creating an environment and forming (proactive) habits that facilitate processes to achieve our goals. Such control is an integral part of the view that implicit attitudes can be properly regarded as part of an agent's character that in every respect is

part. Foucault's idea of practices increases the ways that the individual can be constituted in and through culture, (<http://changingminds.org/explanations/identity/foucault_identity.htm>) (following Foucault and Identity 2020).

legitimately subject to moral evaluation (following Holroyd and Kelly 2016: 130). There is harmony between these two approaches in terms of agent responsibility and at the point of goal setting the proactive engagement of implicit processes to achieve explicit goals.

The model of implicit bias taken forward into Part III will be interactive as described. This credible mainstream approach is in harmony with Dual Process/System models and is proportionate with semicompatibilism in terms of its importance and credibility. While it is principally the interactive model that will be taken forward, there is, as noted, agreement between this model and the approach of Holroyd and Kelly in terms of agent responsibility. Regarding *responsibility* for actions that have as their origin implicit bias, Holroyd and Kelly's conclusion is clear; ' … there is an important sense in which individuals have control over implicit biases' (2016: 130) and with control comes responsibility. While greater understanding is necessary concerning the reflective and impulsive processes that mediate between implicit social cognition and explicit behaviour within the interactive model (Deutsch and Strack 2010: 73), there is no doubt that control in an important sense, and so responsibility, is intrinsic to this model.

5.3 Summary of Part II

Chapter 4 describes implicit bias and the IAT, leading naturally to a description of Dual Process and Dual System models of cognition. If implicit bias is a System One nonconscious associative process, then control or override by System Two conscious rationality appears problematic and without control, responsibility is also problematic.

Chapter 5 considered essentially two approaches to implicit bias and responsibility with the aim of finalising a position on implicit bias. An interactive model was described based on the work of Deutsch and Strack (2010) together with the contrasting approach of Holroyd and Kelly (2016), an approach in accord with Deutsch and Strack's model; importantly, both approaches express individual responsibility for overt behaviour having implicit cognition as its source.[169]

[169] Encouragement for this approach can be found in *An Introduction to Implicit Bias* (G. M. Johnson 2020), where it is noted that 'As methods for studying bias become more sophisticated, so too does our understanding of how bias operates in the minds of individuals. Given the variety, readers might be sceptical that there is even a unified phenomenon to be studied under the heading of implicit bias research. If this is right, it would explain why some data surrounding implicit bias operation just can't be

Before leaving Part II, some broader comments on the idea mentioned above under the heading ecological control, whereby features of the environment can be managed to intentionally influence cognitive processes in particular ways, including automatic features of our mental processes, so allowing indirect control and mitigation of implicit bias related attitudes. While 'Implicit biases are typically thought to reside "inside our heads" ' (Ayala-López and Beeghly 2020: 215) implicit or explicit biases may be described in terms of absorbed controlling images and ideas that originate from within *social structures* that are made manifest by agents (Ayala-López and Beeghly 2020: 216). While within this Thesis there is emphasis on 'psychological' bias, it is *essential* to confirm the importance of structural and environmental factors cannot be overstated. Structuralist accounts position external factors such as inequality as the *cause* of explicit and implicit bias not the outcome. Clearly, such a view has vital implications for bias mitigation; structural change must precede psychological programs that target individual bias directly (following Ayala-López and Beeghly 2020: 211-232) not least because to

> … endorse an analysis of injustice that prioritizes individuals (and especially their mental states) … we are encouraging theorists to remain at the periphery of social problems, ethically and politically speaking, rather than getting to their core. (Ayala-López and Beeghly 2020: 221)

Ayala-López and Beeghly develop an account of implicit bias, inequality and injustice that includes structural *and* individualistic approaches that motivate a more comprehensive and deeper understanding of social injustice; 'both approaches are necessary to explain what's wrong with injustice, why inequalities occur, and how to transform our world (and ourselves) for the better' (2020: 227).

explained using one, monolithic psychological explanation. Instead, we would need a variety of different theories. The purpose of psychological theorizing around implicit bias, then, would be to search for different explanations, describing in what instances they're apt, investigating what, if anything, unifies them, and, importantly, doing all this while ensuring that such explanations are genuinely explanatory' (Johnson 2020: 35).

Part III

Semicompatibilism and Implicit Bias

Chapter 6

Does Implicit Bias Threaten the Semicompatibilist Position?

> *There are two sources of the metaphysical conundrum of human existence. The first is consciousness; the second is freedom.*
>
> Roger Scruton[170]

6.0 Introduction

Why is semicompatibilism chosen as the free will position to be examined? Recall the summary of Chapter 3, where it was noted that semicompatibilism is the gold standard compatibilist position, capturing our intuitions about agency and offering a rich and plausible model of our place in the world. Semicompatibilism sustains our sense of moral responsibility and status as persons whether it happens to be the case that determinism, or indeterminism, is true. John Martin Fischer's semicompatibilism can accept the consequence argument and accommodate the incompatibilist's claim that determinism rules out the type of freedom that allows choice between truly open and available alternatives (regulative control). Motivated by Frankfurt examples, such regulative freedom is claimed to be unnecessary for moral responsibility; guidance control is sufficient to ensure agent responsibility. Part III examines implicit bias and semicompatibilism, seeking to understand if an agent's implicit bias related behaviour is subject to guidance control and therefore performed freely and responsibly, as those terms have been defined within the semicompatibilist model. As noted, on the semicompatibilist view, metaphysical problems arising from an insistence on regulative control (usually as a condition for individual responsibility) together with the truth of causal determinism are avoided. Guidance control gives the agent the kind of freedom

[170] *Modern Philosophy, An Introduction and Survey* (Scruton 2012: 227).

sufficient for responsibility whether or not determinism is true. If issuing behaviour of implicit bias is shown *not* to be subject to guidance control, then the agent does not act freely and responsibly as articulated by the semicompatibilist model. If issuing behaviour is not subject to guidance control *and yet* there are sound and compelling arguments for agent responsibility for behaviour issuing from implicit bias, as described in the previous chapter, then there is surely a deficiency or problem with the semicompatibilist model. If this is the case, it may be possible to suggest amendments or revisions to the semicompatibilist model to take account of the issues brought to light by implicit bias. This is at the heart of the research within this Thesis and a contribution to knowledge within this field. Essentially, Chapter 6 is in two parts. The first and major part examines implicit bias and guidance control, the second part looks at semicompatibilism, luck and implicit bias. Luck is perhaps *the* major threat to compatibilism. A defence of semicompatibilism from the luck problem will be examined in light of implicit bias. Implicit bias appears to be a paradigm example of a source of behaviour that is formed and often continually reinforced by factors that are subject to luck within an agent's behaviour issuing mechanism.

In this chapter I will show the semicompatibilist position on free will and responsibility, developed by John Martin Fischer, is not threatened by the challenge of implicit bias. Implicit bias related behaviour is shown to be subject to guidance control and so agent responsibility in harmony with the models of implicit bias developed in Part II. I will also argue that Cyr's defence of semicompatibilism from the luck problem is not affected when considering the interactive model of implicit bias (Fig 5.1) where implicit attitudes are included within an agent's deliberative standpoint. However, compatibilism is still threatened by the luck problem in the context of implicit bias as characterised by Levy.

6.1 Does Implicit Bias Threaten the Semicompatibilist Position?

Part I presented an overview of free will, including a description of semicompatibilism. Part II considered implicit bias, arriving at a particular characterisation described as interactive Dual System, developed in detail by Deutsch and Strack, *Building Blocks of Social Behaviour* (2010). The essential objective so far has been to understand these concepts, one almost as old as philosophy itself, the other relatively modern. To begin Part III the central idea of guidance control will be explored considering implicit bias.

6.1.1 Implicit Bias and Guidance Control

Given the model of implicit bias developed in Part II, is implicit bias related behaviour subject to guidance control and so appropriate for moral appraisal? Consider the four necessary[171] aspects of guidance control[172] outlined below:

1. An agent has guidance control in so far as *their* deliberation mechanism is appropriately responsive to reasons (Fischer and Ravizza 1998: 41-46).
2. The appropriate way is moderate reasons-responsive (Chapter 3, page 72):
   An agent's responsibility relevant mechanism *K* is moderately reasons-responsive iff:
   a. *K* is regularly *receptive* to reasons, some of which are moral. This requires;
      i. That holding fixed the operation of a *K*-type mechanism, the agent would recognize reasons in such a way as to give rise to an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs.
      ii. That some of the reasons mentioned in (2.a) are moral reasons.
   b. *K* is at least weakly *reactive* to reasons; this requires that the agent would react to at least one sufficient reason to do otherwise, (in some possible scenario),

[171] Todd Long's Paper *Moderate Reasons-Responsiveness, Moral Responsibility, and Manipulation* (2004) was very helpful in preparing this summary.

[172] John Martin Fischer, *Deep Control: Essays on Free Will and Value,* makes a clear summary of semicompatibilism: 'Semicompatibilism is the view that causal determinism is compatible with moral responsibility, quite apart from whether causal determinism rules out regulative control. (Semicompatibilism is officially agnostic about whether causal determinism does rule out regulative control.) Thus, a Semicompatibilist might accept the conclusion of the Consequence Argument, but still hold that causal determinism is compatible with moral responsibility. A Semicompatibilist can thus accommodate a kernel insight of the incompatibilist but also embrace the attractive features of compatibilism, most notably, the *resiliency* of our fundamental views of ourselves (with respect to certain abstract scientific theories). Of course, a Semicompatibilist *need not* accept the conclusion of the Consequence Argument. It is no part of the essence of Semicompatibilism that causal determinism is incompatible with regulative control; rather, the fundamental idea is that moral responsibility depends on how the actual sequence unfolds, not on whether the agent has access to alternative possibilities. Semicompatibilism is, as I have emphasized, an "actual-sequence" model of moral responsibility' (Fischer 2015a: 17).

although it does not follow that the agent could have responded differently to the actual reasons.

   c. *K* is the agent's own; being the agent's own means 'taking responsibility' for *K*. This requires that the agent;
      i. Sees herself as the source of her behavior (which follows from the operation of *K*).
      ii. Believes that she is an apt *candidate* for the reactive attitudes that result from how she exercises her agency in certain contexts.
      iii. Views herself as an agent with respect to (c.i) and (c.ii) based on her evidence for these beliefs (1998: 62-91, 243-244).

3. To be morally responsible is to be an apt candidate for reactive attitudes (1998: 7).
4. An epistemic condition; to be praiseworthy or blameworthy an individual must know, or be reasonably expected to know, what they are doing (1998: 12).

These are the necessary and sufficient ingredients for an agent to affirm guidance control. With focus on the integrative model of implicit bias, I will consider the question carefully, exploring implicit bias and semicompatibilism step by step. Before looking at the elements of guidance control, an appropriate starting point is John Martin Fischer's concluding remarks from *Four Views on Free Will*:

> When we walk down the path of life with courage, or resilience, or compassion, we might not (for all we know) make a certain sort of difference, but we *do* make a distinctive kind of statement. For the semicompatibilist, the basis of our moral responsibility is not selection in the Garden of Forking Paths, but self-expression in writing the narrative of our lives: it is not that we make a difference, but that we make a statement. In writing the stories of our lives, we connect the dots in a way that gives our lives a signature kind of meaning. Even if the name is unexciting, the idea is beautiful. (2007: 82)

John Martin Fischer is clearly concerned with issues that are broader than those explored in this chapter, relating his notion of freedom to questions about the meaning and value of life, rather than focusing on the specific issue of responsibility. I have included this quotation because some of the analysis of semicompatibilism and implicit bias will be detailed, exacting and in an important sense theoretical. I believe it is good to confirm at the beginning that these issues ultimately concern 'our deepest and most basic views about our agency – our freedom and moral responsibility … ' (Fischer *et al.* 2007: 81). Our views on agency, freedom and moral responsibility have direct consequences for broader considerations of the meaning and value of our life; there are strong links between all these issues.

I will look at each of the characteristics of guidance control beginning with moderate reason-responsiveness.

Moderate Reason-Responsiveness

There are three aspects of moderate reason-responsiveness, a, b and c. Each aspect and its subsections, i, ii and iii will be examined, described, for example, as (2.a), or (2.a.i). Then, item 3 moral responsibility and the reactive attitudes, and finally item 4 the epistemic condition will be examined. Additional subsections will be needed and explained as necessary.

To begin then, with item (2.a) *K* is regularly *receptive* to reasons, some of which are moral.

1. An agent has guidance control as far as their deliberation mechanism is appropriately responsive to reasons.

2. The appropriate way is moderate reasons-responsive:

   An agent's responsibility relevant mechanism *K* is moderately reasons-responsive iff:

   a. *K* is regularly *receptive* to reasons, some of which are moral. This requires;

      i. That holding fixed the operation of a *K*-type mechanism, the agent would recognize reasons in such a way as to give rise to an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs.

      ii. That some of the reasons mentioned in (2.a) are moral reasons.

The essential question is, considering the integrative model of implicit cognition, is it the case that an agent is receptive to reasons in the sense described under (2.a)? From the outset, this is not a straightforward question. The integrative model includes a behavioural decision-making component that is receptive to reasons and so plausibly leads to responsible behaviour, as previously discussed. However, it is not immediately clear how implicit bias related behaviour could form an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs as item (2.a.i) requires. Such an understanding would clearly depend on how well the third party understood the agent, but it seems initially *un*reasonable to expect a third party to be able to fully understand an agent's implicit biases and their effect on behaviour considering an agent's likely claims of contrary values and beliefs. A third party would likely see *inconsistency* between an agent's explicit claims and their implicit bias influenced behaviour, not an *overall* understandable pattern.

To reach a position on this question I will look at John Martin Fischer's *Responsibility and Control: A Theory of Moral Responsibility* (1998: 71-73). Here, the role of the 'third party' is described in more detail. An imaginary interview is conducted by a third party where the agent is asked to give reasons for their actions in actual and hypothetical scenarios. The third party considers this information and their background knowledge of the agent, seeking to understand if there is a pattern in their set of reason recognition. John Martin Fischer describes regular reasons-receptivity as 'reasons-receptivity that gives rise to a minimally comprehensible pattern, judged from some perspective that takes into account subjective features of the agent, (i.e., the agent's preferences, values, and beliefs), but is also not simply the agent's point of view … and is minimally grounded in reality' (1998: 71-73). The position so far is that for implicit bias issuing behaviour there is an *understandable pattern*, minimally grounded in reality, in the sense that a third party, not the agent, could see patterns of consistent behaviour even if there is inconsistency between agent reported values and beliefs and those reflected in actual behaviour, (in particular circumstances). Consistency of outcome is present (with stress level and cognitive load fixed) within the integrative model of implicit bias, there is regularity, as essentially required for guidance control.

However, while it is possible that a third party could recognise a pattern of agent behaviour, condition 2.a.i requires the *agent* to recognize reasons in such a way as to *give rise* to an understandable pattern. Is it possible for an agent to recognise reasons in the case of implicit bias? While a third party may recognise a pattern in behaviour, inconsistent with declared values, if the agent is unaware of (is not receptive to) reasons for such behaviour then John Martin Fischer's condition is not fully satisfied. Consider the following example. If someone has an implicit bias against a particular racial group and as a result does not employ members of that group, yet when this is brought to their attention, they take steps to correct their behaviour, then it can be plausibly claimed the agent is responsive (receptive and reactive) to reasons. Their behaviour may also be part of a recognisable pattern in the sense that whenever their bias is highlighted, they take corrective steps. However, in this example, being receptive to reasons, crucially, is not taking place at the time of the biased behaviour. If an agent is not receptive to reasons at the time the biased behaviour takes place and is not alerted later when receptivity to reasons (provided by a work colleague perhaps) *may* be active, in such cases, it seems the agent cannot be said to be responsive to reasons (receptive or reactive) and therefore cannot assert guidance control or be responsible.[173]

John Martin Fischer responds to problems of this form by invoking the notion of nonreflective behaviour (1998: 85) and the tracing principle. Nonreflective behaviour relates to actions for which agents are responsible that do not issue from an exercise of practical reason, the capacity for deciding what to do by reflection and deliberation. Examples are given of nonreflective behaviour, such as a woman who automatically holds the door open for a stranger entering after her; ' … actions that arguably proceed not from deliberation (and thus not from a mechanism of practical reason), but rather from something like habit, character, or instinct' (1998: 85). Moderate reason-responsiveness seems to need deliberation; being receptive and reactive to reasons looks very much like an activity requiring reflection and deliberation. However, John Martin Fischer does not want cases of *nonreflective* behaviour to be excluded from guidance control and responsibility, as this would be, in John Martin Fischer's view, an unacceptable, (and incorrect), limitation of the semicompatibilist position. It is implausible to suggest that all nonreflective actions are unworthy of praise or blame, that a lifeguard is in a moral sense not responsible and praiseworthy for a spontaneous and dangerous rescue. The integrative model of mechanisms such as attitudes, stereotypes and prejudice that mediate the influence of implicit social cognition on behaviour (see Fig 5.1) *has* a component of behavioural decision making and associated reflection. Therefore, the integrative model of implicit bias does allow the possibility of reflection

[173] I am grateful to my Supervisor Dr. Nash for highlighting this issue and suggesting the illustrative example. As discussed within this Thesis, whether implicit biases respond to reasons is far from universally agreed. Under high stress deliberation would certainly be impaired.

and so is not entirely subject to John Martin Fischer's concerns unless the agent is under high stress and/or cognitive load. Under cognitive load, (during an IAT for example), or a particularly emotional construal of a situation, issuing behaviour may be very close to an impulsive response with increasing absence of reflection. It is *possible* that reflective behaviour within the chosen model of implicit bias could become compromised or even impossible. It is important therefore to understand if and how moderate reason-responsiveness could function if conscious reflection became impossible. Importantly, John Martin Fischer responds, that nowhere in the description of moderate reason-responsiveness is it required that this kind of mechanism be 'practical reasoning' (1998: 86) and offers examples to show why 'practical reasoning' is unnecessary.

The first example describes a driver who, as a matter of habit and without deliberation, always takes a particular turning to arrive at their chosen destination. Surely, there is no absence of responsibility in this scenario. In the next example the usual turning is blocked and the driver, again without deliberation, takes the next available turn and continues their journey to the same destination by a slightly longer route. Responsibility is assessed by holding fixed the relevant mechanism, the actual-sequence mechanism, and asking whether there are possible scenarios in which certain things are different and the agent acts differently. In the second example certain things *are* different, the agent *acts* differently *and* the agent's actual sequence that issues in action is from a *non*reflective mechanism. John Martin Fischer suggests this helps to show that such a *non*reflective mechanism is very plausibly still moderately reasons-responsive (1998: 86); an alternative route was taken to the desired destination.

So, does the requirement of reasons recognition rule out moral responsibility for actions that issue from a clearly nonreflective mechanism? John Martin Fischer claims it does not, arguing essentially that recognition of reasons does not entail that an agent deliberates on such reasons in terms of what to do, for example, how to solve a problem such as a blocked turning when driving towards a place of work. Reasons-recognition and a nonreflective mechanism are compatible and the agent retains moral responsibility for actions that issue from an explicitly nonreflective mechanism (1998: 87, Fischer and Tognazzini 2009).

If it is assumed that nonreflective behaviour is such that even *recognition* of reasons is ruled out, then moderate reason-responsiveness of the actual operative mechanism would likewise be ruled out. However, John Martin Fischer ingeniously responds to this difficulty by invoking the tracing principle whereby moral responsibility may exist now due to guidance control operating at some appropriate point in the *past*, prior to the action, when the reason for the action was formed. The general principle may be shown by example; a driver, so intoxicated that reason recognition is impossible, had at an earlier time active guidance control when the decision to begin drinking was made, while knowing that driving would take place later. The driver is responsible for an accident now, even though not moderately reason-responsive, due to behaviour in the past when guidance control *was* active (1998: 89).[174] John Martin Fischer acknowledges this is not an exhaustive response to the 'problem' that moderate reason-responsiveness must be receptive and reactive to reasons; given a 'strong' understanding of nonreflective behaviour, where the mechanism that produced it is incompatible even with the *recognition* of reasons, there are still cases where agents are plausibly responsible.

It has been important to consider guidance control, particularly moderate reason-responsiveness, in light of the possibility of nonreflective behaviour because although the adopted integrative model of implicit bias includes reflective processing, as mentioned, under pressure this could be compromised; 'straining the capacities of reflective processing decreases self-control … ' (Deutsch and Strack 2010: 71) and in the limit eliminates it completely. John Martin Fischer shows that reason receptivity, the vital ingredient of moderate reason-responsiveness, is not adversely affected by nonreflective behaviour; the issuing mechanism is still the same. The 'stronger' interpretation of nonreflective behaviour is not completely resolved, i.e., when *recognition* of any reasons is

[174] A fictional, but plausible, example of a life or death outcome that essentially depends on acceptance or rejection of a philosophical position or claim, in this case the issue of 'tracing', (responsibility for an action now, although an agent is not currently reason-responsive due to behaviour in the past when control was active), is provided by Series 9, Episode 7 of the T.V. show Law and Order: Special Victims Unit, *Blinded*, first shown on November 13th 2007. In this episode the contentious issue is the nature of the guilt of a 'murderer' who has committed several crimes due to an earlier decision, when considered to be sane, to stop taking behaviour controlling prescribed medication. The position taken on the strength or legitimacy of the tracing principle in this example has life or death implications; criminal responsibility for recent actions because of the earlier decision to quite medication would probably lead to trial and execution, alternatively, if it is decided that the murderer acted while lacking responsibility *at the time* of the crime when not in control due to mental illness and earlier rational decisions were irrelevant, then the outcome would be transfer to a psychiatric hospital.

completely ruled out, but tracing, as described, suggests one way of incorporating such cases within the guidance control model.

The discussion above is motivated by the question, can implicit bias provide a reason to act? This question is clearly important because it is condition 2.a.i of guidance control; an agent must recognize *reasons* in such a way as to give rise to an understandable pattern of behaviour. John Martin Fischer shows that reason receptivity, is not adversely affected by nonreflective behaviour. *If* implicit bias is considered as the source of nonreflective behaviour, then John Martin Fischer's arguments are not affected. Requirement 2.a.i of guidance control, (that an agent would recognize reasons in such a way as to give rise to an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs), has been challenging to analyse considering implicit bias. I believe it is reasonable and defensible to conclude this requirement of guidance control, and hence responsibility, *is* satisfied (while noting some limitations of tracing) when behaviour is initiated by implicit bias within the operating mechanism. To sum up, John Martin Fischer's semicompatibilism accommodates situations that have a nonreflective mechanism, but complete absence of receptivity to reasons is, (not surprisingly), problematic. In the case of implicit bias influenced actions, in a stressful situation, there may only be the possibility of minimal reflection. However, for an agent to act in ways that reflect biases there must be *some* reason that initiates such behaviour, even though the reason is not consciously recognised. In this sense the agent must be receptive to reasons and importantly, semicompatibilism accommodates this form of nonconscious reason receptivity and behaviour mechanism. In the earlier example, the driver's action and awareness of the reason to act may not be conscious, but the action must be the result of being *receptive* to a reason (the blocked exist), otherwise no action would take place.

It has been argued (Holroyd 2012: 295) that implicit biases are responsive to (bad) reasons in the sense that implicit associations are formed in response to, for example, pervasive stereotypes; their existence is in response to reasons (even if they are marginal or incorrect reasons). While arguably responsive to reasons in their formation, implicit bias considered as an automatic process also responds to reasons when manifest in action. Similar examples to those above are given supporting this claim, examples of automatic processes, (such as playing a good tennis shot under pressure), that are clearly responsive to reasons and lead to responsible action. The form of indirect control over implicit bias manifestation, (ecological control), described earlier, is an example of managing the environment to make the agent (more) responsive to good reasons.

2. The appropriate way is moderate reasons-responsive:
An agent's responsibility relevant mechanism K is moderately reasons-responsive iff:

  a. K is regularly receptive to reasons, some of which are moral. This requires;

    ii. That some of the reasons mentioned in (2.a) are moral reasons.

Exploration of the three aspects of moderate reason-responsiveness, (2.a, 2.b and 2.c), continues, looking at item (2.a.ii), that some of the reasons mentioned in (2.a) are moral reasons. Implicit bias related behaviour is clearly charged with moral content. Implicit (or explicit) attitudes toward particular social groups, for example, frequently include morally relevant qualities such as untrustworthy, cowardly and so on (often referred to as prejudices), (Eberhardt 2019: 31). It seems the requirement of guidance control, that an agent recognizes reasons some of which are moral reasons, is straight forwardly satisfied in the case of implicit bias and related behaviour. Implicit biases very often have at their heart negative moral evaluation of particular groups, see for example Jennifer Eberhardt (2019: 23). There will be more to say about 'some of the reasons mentioned in (2.a) are moral reasons' later (page 179).

2. The appropriate way is moderate reasons-responsive (Chapter 3, page 72):
   An agent's responsibility relevant mechanism K is moderately reasons-responsive iff:
      b. K is at least weakly reactive to reasons; this requires that the agent would react to at least one sufficient reason to do otherwise, (in some possible scenario), although it does not follow that the agent could have responded differently to the actual reasons.

K is at least weakly reactive to reasons; this requires that the agent would react to at least one sufficient reason to do otherwise, (in some possible scenario), although it does not follow that the agent could have responded differently to the actual reasons.

There are three conditions associated with *strong* reasons-responsiveness: (i) The agent is strongly *receptive* to reasons, where receptive means the capacity of an agent to recognize the reasons that exist. (ii) The agent is strongly *reactive* to reasons, where reactive means an agent has the capacity to translate reasons into choices, (and then subsequent behaviour). (iii) The agent produces actions that are in accord with choice. A mechanism moves from being strongly reasons-responsive to being *weakly* reasons-responsive (or not responsive at all) because of a 'deficiency' in any of these three areas.

Reactivity to reasons and receptivity to reasons that constitute the responsiveness relevant to guidance control and moral responsibility are asymmetric; a very weak sort of reactivity is all that is required, whereas, a stronger sort of receptivity to reasons is necessary for this kind of responsiveness (Fischer and Ravizza 1998: 69). To be (very) weakly reactive to reason, the agent (when acting from the relevant mechanism, the actual-sequence mechanism that leads to the action), must simply display *some* reactivity. Some reactivity gives plausibility to the agent's actual mechanism having 'executive power' to react to the actual incentive to do otherwise, but note, there is not a demand for regularity with respect to reactivity and there is not an *explicit* requirement for *reactivity* to moral reasons (1998: 79). There is a further point; receptive and reactive aspects of reason responsiveness are a general capacity of the agent's *mechanism*, rather than a particular ability of the agent in the sense of possessing alternative possibilities - the freedom to choose and do otherwise (1998: 75).[175]

How does this complex aspect of guidance control relate to implicit bias characterised by the integrative model? Specifically, is implicit bias, reason responsive in the sense that issuing behaviour is weakly reactive to reason? The answer appears to be yes, as straightforward and plausible examples can be given of an agent whose implicit bias related behaviour is weakly reactive to reasons. The actual mechanism issues in the agent doing *A* in the actual world; on being told by telephone that intruders have entered their home, acting impulsively, the agent under stress takes a lift home from a young rather than elderly driver because of implicit bias relating to elderly drivers based on a stereotype of older drivers as indecisive and making only protracted progress when fast progress is needed. However, 'there is some possible world with the same laws of nature in which a mechanism of this kind is operative in this agent and there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise

[175] The preceding description in this section follows closely John Martin Fischer as cited.

for this reason' (Fischer 2015b: 188). Continuing with this example, with the small capacity for deliberation available, the agent in some possible world remembers that younger drivers may not have sufficient experience to make safe judgements when driving under *pressure* and so chooses what could be a slightly slower but more assured journey home to confront the intruders. This example turns on the agent recognizing this reason in some possible world. If the level of stress is such that only impulsive reactions are possible then recognizing reasons is impossible, but recall John Martin Fischer's argument described above, that nonreflective recognition of reasons (and reaction to reasons) is accepted and does not entail that an agent deliberates on such reasons in terms of what to do.

To summarise, when the requirements for moderate reason-responsiveness, that *K* is regularly *receptive* to reasons, (some of which are moral), and *K* is weakly *reactive* to reasons, are considered in the context of implicit bias, I believe it is reasonable to claim that issuing behaviour of implicit bias *is* moderately reason-responsiveness in terms of these requirements, supporting the view that an agent is morally responsible for their actions. A conclusion in accord with the integrative model of implicit bias.

2.c. K is the agent's own; being the agent's own means 'taking responsibility' for K. This requires that the agent;

i. Sees herself as the source of her behaviour (which follows from the operation of K).

ii. Believes that she is an apt candidate for the reactive attitudes that result from how she exercises her agency in certain contexts.

iii. Views herself as an agent with respect to (2.c.i) and (2.c.ii) based on her evidence for these beliefs.

(2.c.i) Sees herself as the source of her behaviour (which follows from the operation of K). Item (2.c) is the third of the general requirements for the mechanism that issues in behaviour to be reason-responsive in the appropriate sense, and is the most challenging to analyse: 'Taking responsibility' is one of the ways that a mechanism that leads to action becomes the agent's own (Fischer and Ravizza 1998: 207). At first sight, items (i – iii) appear completely at variance with the pervasive idea that an agent is unaware of their implicit bias related behaviour. On this popular account, the possibility of seeing oneself as the source of such behaviour (2.c.i) appears remote.[176] However, recall that John Martin Fischer during discussion of nonreflective behaviour (page 155) is keen to show that the semicompatibilist/guidance-control model accepts that an agent can take responsibility in an implicit, nondeliberative way:

> This process (taking responsibility) may involve conscious and deliberate reflection, but it need not. Just as a person who acts for a reason need not explicitly formulate the reason or consciously invoke it as an action guide, so a person can take responsibility in an implicit, nondeliberative way. (Fischer and Ravizza 1998: 214)

While this is helpful in supporting the view that agents may still be responsible for actions that are conducted in a nondeliberative way, for example, (following the popular view), actions of an agent unaware of their implicit bias related behaviour, it does not get to the heart of the issue, that of *seeing herself as* the *source* of her implicit bias issuing behaviour. Where 'source' is the perception of *oneself* as the cause of events in the world, originating from desires, beliefs and intentions, not the result of an accident or the intentions and actions of others. Drawing on the broader characterisation of implicit bias described during discussion of Holroyd and Kelly (2016) it seems reasonable to suggest that the requirement of 'taking responsibility' by seeing ourselves as the source of our behaviour appears sufficiently close to Holroyd and Kelly's description of implicit attitudes as part of who the agent is, part of character and *source of behaviour*, which as a whole is subject to moral evaluation.[177] I am suggesting the requirement of guidance control, that an agent takes responsibility by *seeing themselves as the source of their behaviour*,

[176] Implicit attitudes and awareness are discussed later in this chapter. Two particularly relevant Papers are *Implicit Attitudes and Awareness* (Berger 2018) and *Are 'Implicit' Attitudes Unconscious?* (Gawronski, Hofmann, and Wilbur 2006).

[177] One of the four necessary aspects of guidance control is the epistemic condition, discussed later in this Chapter. There are, as would be expected, similarities between the epistemic condition relating to guidance control and ecological control; 'whether an agent can exercise ecological control, (a form of control that individuals have with respect to implicit bias), depends on whether she is aware of these possibilities (and aware of the phenomena of implicit bias, and that she may be affected by it). The mere possibility of having ecological control is not sufficient for implicit biases to be considered as 'part of the agent' and hence morally evaluable. In addition to the control conditions, epistemic conditions must also be met as well' (Holroyd and Kelly 2016: 127).

is satisfied in the case of implicit bias related behaviour when, (following Holroyd and Kelly), implicit attitudes are considered part of who the agent is, part of character *and a source of behaviour* which as a whole is subject to moral evaluation. However, there are issues with this approach. First, the requirement for guidance control is to *see herself* as the source of her behaviour and this has not yet been addressed. More troubling, there is a sense of circularity. Recall the argument from Chapter 5 (page 136):

1. *If* an agent has the relevant control and responsibility for implicit bias, *then* implicit bias reflects an agent's character.
2. Individuals *do* have relevant control/responsibility for implicit bias.
3. Therefore, implicit bias reflects an agent's character, (and can legitimately be morally appraised).

That implicit bias reflects an agent's character is a conclusion based on premises concerning *control*, that individuals do have relevant control/responsibility for implicit bias. I then use the conclusion that implicit bias reflects an agent's character (and a source of behaviour) to make claims *about* a type of control, i.e., guidance control; this is an unsatisfactory approach.

Consider the question from a different perspective; the popular idea of the nonconscious nature of implicit bias and issuing actions has much counter empirical evidence and supporting scholarship. On the basis of such evidence the model of implicit bias employed to critique semicompatibilism and guidance control contains *reflective behaviour decisions*.[178] Actions that result from individual *control and reflective decision* making seem quite naturally 'assignable to individual agents as sources' (Nagel 2003: 229) and this is the natural internal sense of our own freedom and agency. With control and reflective decision making[179] there must surely be a sense of being the source of issuing

[178] It may be helpful at this point to reiterate the overall objective of the Thesis. The chosen implicit bias model includes the conclusion that agents *are* responsible for issuing behaviour. Does this model satisfy the conditions for guidance control and so responsibility? Both approaches *should* reach the same conclusion, if this is not the case then some form of amendment or further consideration is necessary with respect to the semicompatibilist/guidance-control model, (or the implicit bias model, or both, but it is semicompatibilism that is the target here, with the chosen characterisation of implicit bias held fixed, as reflected in the title of the Thesis).

[179] See also later discussion of the epistemic condition for responsibility.

behaviour. To make reflective behaviour decisions is surely to see yourself *as the source of* such decisions. While seeing oneself as the source of implicit bias related behaviour is contrary to the popular conception of implicit bias as essentially unconscious, I believe that the chosen model, that includes an element of reflective decision making and control, allows the agent to take responsibility by 'allowing an agent to see herself as the source of her behaviour'. The issue of awareness of implicit bias is a constant theme throughout Part II and III; the position settled on is that agents are aware of and responsible for their implicit biases and related behaviour, *but* such awareness and responsibility is not binary but expressed in degrees. There is further discussion of awareness shortly, at the beginning of the discussion of the third and final requirement of 'taking responsibility' (2.c.iii).

To close this section (2.c.i), significantly, Brownstein comments that empirical literature is quite mixed concerning conclusions about awareness of implicit attitudes and whether their influence can be controlled (2016a: 770), but after reviewing the empirical data Brownstein concludes that agent awareness of the content of implicit attitudes is often the case, but awareness of the influence of implicit attitudes on behaviour is less so. However, it is *clear* from Brownstein's comments that agent awareness of the influence of implicit attitudes on behaviour is possible, but there is some uncertainty over the *extent* of awareness. Brownstein points out that agent awareness of the influence of their *ex*plicit attitudes on behaviour is equally problematic, (see also Gawronski, *Six Lessons for a Cogent Science of Implicit Bias and Its Criticism* (2019: 578) where a similar point is made).

Examination so far has shown that from two different perspectives, (guidance control and the integrated model), implicit bias and issuing behaviour are subject to legitimate moral appraisal; the nature of implicit bias, as characterised, is subject to guidance control, the freedom and responsibility defining condition, not subject to the truth or falsity of determinism or the exercise of regulative control.

Next, I will look at the second condition for 'taking responsibility'; an agent's belief that she is an apt candidate for the reactive attitudes that result from how she exercises her agency in certain contexts, (2.c.ii).

2.c. K is the agent's own; being the agent's own means 'taking responsibility' for K. This requires that the agent;

ii. Believes that she is an apt candidate for the reactive attitudes that result from how she exercises her agency in certain contexts.

The model of implicit bias developed in Chapter 5 does not explicitly consider the reactive attitudes. However, I do not see dissonance between an agent's belief that she is an apt candidate for the reactive attitudes that result from how she exercises her agency in certain contexts and ecological control, a key component of Holroyd and Kelly's (2016) approach and an important element in the characterisation of implicit bias. It is reasonable then to construe the reactive attitudes as potential props, (to use Holroyd and Kelly's expression), part of an individual's wider strategy to take ecological control. This is tentative, but in terms of the second condition for 'taking responsibility', that an agent should believe that she is an apt candidate for the reactive attitudes, no obvious or significant omissions or problems with semicompatibilism or guidance control are revealed in light of implicit bias as it has been characterised.

The third and final requirement of 'taking responsibility' (2.c.iii).

2.c. K is the agent's own; being the agent's own means 'taking responsibility' for K. This requires that the agent;

iii. Views herself as an agent with respect to (2.c.i) and (2.c.ii) based on her evidence for these beliefs.

Before addressing the third and final requirement of 'taking responsibility' I will make some further comments on agent awareness of implicit attitudes and offer further support for the chosen model of implicit bias. This is beneficial before considering an agent's awareness of themselves *as* an agent and taking responsibility for behaviour. A robust position on implicit bias is necessary in order that the critique of semicompatibilism be as compelling and sound as possible. The proposed view of implicit bias, based on Holroyd and Kelly (2016) and Deutsch and Strack (2010), concludes that agents *are* essentially responsible for issuing behaviour, based essentially on considerations of character from Holroyd and Kelly and the possibility of reflective behavioural decision making from the Deutsch and Strack perspective. At an individual level, a situation can be imagined, where, after completing an IAT the candidate comments, 'I am not aware of this, surely there must be a mistake, my beliefs are not like this at all.' It is the stark contrast between 'Sees herself as the source of her behaviour' (2.c.i) and the surprise and disbelief experienced by many after completing an IAT and exposure of 'hidden' attitudes contrary to explicit beliefs that calls for further mention before returning to complete consideration of 'taking responsibility' item (2.c.iii).

The pervasive popular view concerning explicit and implicit attitudes is that agents are *aware* of explicit attitudes, which are *conscious* and unaware of unconscious implicit attitudes. This provides a fairly straightforward explanation of the observation that agents 'readily articulate their explicit attitudes, but rarely if ever report - and often deny - their revealed implicit ones. A natural explanation of this observation is that individuals are aware of the former but *not* aware of the latter' (Berger 2018: 2). There is, of course, a counter position to this view, expressed within a significant number of Papers, for example, the excellent *Are 'Implicit' Attitudes Unconscious?* (Gawronski, Hofmann, and Wilbur 2006). I will look at this Paper together with *What is Implicit Bias?* (Holroyd, Scaife, and Stafford 2017), adding clarity to the conscious/nonconscious distinction, plus generally improving intelligibility and defence of the chosen model of implicit bias.

Holroyd, Scaife and Stafford (2017), (henceforth Holroyd), ask what is 'implicit' and what is 'bias'? The form the answer takes depends on the ambition of the enquirer. An account of implicit bias hopes to achieve one or several of the following desiderata;

D1: Distinguish implicit from explicit mental states or processes.

D2: Capture interesting cases of dissonance between agents' professed values and the cognitions driving responses to these measures.

D3: Formulate interventions for changing bias or blocking discriminatory outcomes.

D4: Accommodate or explain the full range of the phenomena captured by indirect measures.

D5: Gain traction in addressing problems of marginalisation and under representation, and draw attention to complicity in these problems (Following Holroyd and others 2017).

The task now is to look more carefully at implicit bias in terms of unconscious implicit attitudes and conscious explicit attitudes. Using unconscious and conscious attitudes to distinguish implicit from explicit mental states or processes (D1) appears to run well with the observation previously mentioned that many find the outcome of indirect measures such as the IAT surprising, even shocking, something that did not make itself known to conscious awareness (D2). But what is the agent *unconscious of?* There are at least three aspects of an attitude that could be considered unconscious (Gawronski and others 2006: 487), (Holroyd and others 2017: 4): (i) Source awareness - an agent is unaware of the cause of an attitude towards an object. Gawronski points out, 'both self-reported *and* indirectly assessed attitudes may be characterized by a lack of source awareness (2006: 489). (ii) Content awareness - an agent is unaware of the attitude itself. Low correlation between self-reports and indirectly assessed attitudes is often suggested as *empirical evidence* for an unconscious characterisation of implicit bias. (iii) Impact awareness - an agent is unaware of the influence a given attitude has on other psychological processes. Considering only item (ii) content awareness, Gawronski says,

> in contrast to this conclusion, (that participants are generally unable to report certain attitudes because they are unconscious), there is now accumulating evidence that self-reported attitudes are *systematically related to* indirectly assessed attitude. Moreover, the relative size of the correlation between self-reported and indirectly assessed attitudes seems to depend on a variety of different variables related to basic psychological as well as methodological factors. (added emphasis 2006: 489)

On Gawronski's account; (i) self-reported and indirectly assessed attitudes are systematically related, (ii) such findings are in contrast to the widespread assumption that people generally have no conscious access to indirectly assessed attitudes, (iii) it seems that people *are* consciously aware of the attitudes assessed by indirect measures, (iv) whether or not these attitudes are reflected in self-report measures depends on a variety of factors pertaining to cognitive, motivational and methodological variables, (v) it seems that people do have introspective access to their attitudes, as they are reflected in indirect attitude measures. However, these attitudes may not be reflected in self-reported evaluations when cognitive, motivational, or methodological factors undermine their impact on explicit self-reports. (following closely Gawronski and others 2006: 490). This recalls the discussion of Deutsch and Strack, *Building Blocks of Social Behaviour* (2010), where mechanisms that mediate the influence of implicit social cognition (for example, motivation), were shown as a unified system of related elements, (Fig 5.1).

Is the property of being an *unconscious* attitude useful in distinguishing implicit from explicit mental states or processes (D1) and does it capture interesting cases of dissonance between agents' professed values and the cognitions driving responses to these measures (D2)? Holroyd concludes it does not. Based on various sources of empirical data, Holroyd believes 'the evidence suggests […] individuals *have* some awareness of the cognitions revealed on such (indirect) measures' (added emphasis 2017: 4). Essentially, it is argued the notion of implicit as unconscious is not supported by available empirical evidence vis-à-vis the *actual* awareness of agents and does not unequivocally distinguish implicit from explicit states.

The above, I suggest, supplies added justification for the chosen model of implicit bias, in the sense that, having conscious awareness[180] within a model of implicit bias is shown again to have credible theoretical and empirical support. It is clearly important, as previously mentioned, to have a robust model of implicit bias in order that the semicompatibilist position be challenged in a meaningful and defensible way.

Returning to moderate reasons-responsiveness, that an agent's responsibility relevant mechanism *K* is moderately reasons-responsive iff *K* is the agent's own, where being the agent's own means 'taking responsibility' for *K*. This requires, in addition to (2.c.i) and (2.c.ii) previously considered, that an agent views herself as an agent with

[180] There is *much* debate about the nature of awareness. For example, Jacob Berger claims there is 'much evidence that we can be aware of implicit attitudes, it is plausible that we are not aware of them in the same subjectively unmediated way that we are aware of our explicit attitudes. […] An attitude is implicit just in case one is not aware of it in a subjectively unmediated way' (2018: 20). I have not argued for any *specific* conception of awareness within my model of implicit bias, other than an awareness that gives the *possibility* of reflection as shown within Fig 5.1.

respect to (2.c.i) and (2.c.ii) based on evidence for these beliefs (2.c.iii).[181] John Martin Fischer includes this item to address a particular form of manipulation whereby an agent's responsibility is somehow (electronically) implanted. In other words, the individual's view of herself as an agent and an apt candidate for the reactive attitudes is electronically implanted. John Martin Fischer says, quite reasonably, such a view 'is not formed in the *appropriate* way. But the relevant notion of appropriateness must remain unanalyzed' (1998: 236). The comment 'must remain unanalyzed' is puzzling because John Martin Fischer takes time to explain how mechanisms become the agent's own, giving examples *of* the appropriate way that agents can and do take responsibility. What sort of evidence leads an agent to view herself as an agent with respect to (2.c.i) and (2.c.ii)? The personal sense of having ongoing acceptance and *participation* within a moral community that believes responsible agents *are* the source of their behaviour and candidates for the reactive attitudes is perhaps sufficient evidence for the agent. However, this has a sense of circularity and is not entirely convincing.

With respect to the basic question, is issuing behaviour of implicit bias subject to guidance control, where item (2.c.iii) is part of the characterisation of guidance control, there appears no obvious problem for an agent to view themselves *as* an agent with respect to (c.ii), (an apt *candidate* for the reactive attitudes), based on evidence for such a belief. Reactive attitudes in response to our behaviour may be accepted, countered strongly or rejected entirely by an agent, but as a member of a moral community it is accepted that *expression* of reactive attitudes is appropriate and to be expected. Where such behaviour may include implicit bias issuing behaviour. Such a claim may seem unrealistic, in that 'appropriate and to be expected' does not reflect accurately all reactions to praise and particularly blame. However, even when blame is not well received, the right to blame, (if expressed in a reasonable way and without blatant double standards), is not in principle usually contested.

An agent's evidence for believing herself *to be* an agent with respect to (2.c.i), as the source of implicit bias related behaviour, could be gained from measuring, observing or in some way checking the outcome of ecological control. Clearly, an important part

[181] For further discussion see *Reasons-Responsiveness and Ownership-of-Agency: Fischer and Ravizza's Historicist Theory of Responsibility,* Section 3 Evidence Sensitivity in Psychological Development, Sub Section 3.1 In owning one's self-conceptions (Zimmerman 2002: 219).

of working towards reducing/eliminating bias and influenced behaviour when employing ecological control is (if possible) confirming that improvements *are* being achieved. Measurable improvement in implicit bias related behaviour, caeteris paribus, would tend to support the validity of ecological control and provide perhaps supporting evidence for its methods and claims; that implicit attitudes can be properly regarded as part of 'who the agent is', part of character, which is as a whole subject to moral evaluation[182] (Holroyd and Kelly 2016: 130). Taking ecological control and reflecting on evidence of personal achievement in mitigating implicit bias related behaviour surely supports a sense of personal agency, of being in an important sense the source *of* behaviour, (and as a source, an apt candidate for reactive attitudes).

Moral Responsibility and the Reactive Attitudes

Item 3, to be morally responsible is to be an apt candidate for reactive attitudes, has been discussed under item (2.c.ii) above. I will now look at item 4, John Martin Fischer's epistemic condition.

The Epistemic Condition

4. An epistemic condition; to be praiseworthy or blameworthy an individual must know, or be reasonably expected to know, what they are doing (Fischer 1998: 12).

The epistemic condition is essentially the intuitively reasonable and plausible condition that 'an agent is responsible only if he knows the particular facts surrounding his action, and, acts with the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13). Using John Martin Fischer's expression of the epistemic condition, the relevant question becomes, is an agent aware of 'particular facts surrounding an action' and do they act with 'the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13) when actions are influenced by implicit bias? John Martin Fischer makes it clear at the outset

[182] Holroyd and Kelly claim that it is possible for individuals to exercise ecological control over implicit biases. 'But whether, for any individual, they can in fact exercise it depends on whether they are aware of these possibilities (and aware of the phenomena of implicit bias, and that they may be affected by it). So, the mere possibility of having ecological control is not sufficient for implicit biases to be considered as "part of the agent" and hence morally evaluable. In addition to the control conditions, epistemic conditions must also be met as well' (2016: 127). Recall the discussion on page 143.

that the epistemic condition is not something he will focus on, rather, his attention will focus on the freedom-relevant condition for moral responsibility. It may be noted that attention is again on awareness; the first part of the epistemic condition above requires awareness by the agent of particular facts surrounding an action. While John Martin Fischer does not dwell on the epistemic condition, further reading is suggested (1998: 13); Joel Feinberg *The Moral Limits of the Criminal Law Volume 3: Harm to Self* (1989).

Feinberg's Paper is clear and interesting, but does not help to respond to the question, is the agent aware of 'particular facts surrounding an action' and do they act with 'the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13) when actions are directly influenced by implicit bias? To move the investigation forward it is necessary to look at these epistemic conditions in more detail, as outlined below:

(i) Describe the essential features of the epistemic condition debate.
(ii) Consider what John Martin Fischer says about the epistemic condition and guidance control by looking in more detail at 'particular facts surrounding an action' and 'the proper sort of beliefs and intentions'.
(iii) Consider if issuing behaviour of implicit bias satisfies John Martin Fischer's epistemic condition of guidance control and responsibility. The moral sense, and responsibility for an action when completely unaware of its moral significance will also be discussed.

While discussing the above I will draw on the following; *Responsibility: The Epistemic Condition* (Wieland 2017), *Culpable Ignorance* (Smith 1983) and *The Epistemic Condition for Moral Responsibility* (Rudy-Hiller 2018).

(i) Describe the essential features of the epistemic condition debate.
The Stanford Encyclopedia of Philosophy states the epistemic condition is usually considered to have four epistemic requirements: 'awareness of action, moral significance, consequences, and alternatives' (Rudy-Hiller 2018). Briefly expanding each requirement. Awareness of action straightforwardly means that for the agent to be *directly* responsible they must be aware of what they are doing, aware of performing the action under consideration. Using an example from *The Stanford Encyclopedia of Philosophy* (Rudy-Hiller 2018), the 'action under consideration' is John's movement of a switch that he believes

will turn on the light but actually starts a treadmill on which Mary is standing, who falls and is injured; it seems John cannot be blamed for the fall because he is unaware of the action under consideration, the action that unknowingly started the treadmill. If John's ignorance of the actual function of the switch is culpable, for example, because of lack of attention during a safety briefing or some other negligence, then there are compelling arguments for blaming John for Mary's injury. Such consideration opens a substantial debate that seeks to answer general questions such as can ignorance (culpable or otherwise) provide a moral excuse?[183] In the context of implicit bias, questions could be of the form, does ignorance of a tendency to act in a biased way make the action less culpable (or even excused), and what are the implications for blame and responsibility?

The second epistemic requirement for direct responsibility, awareness of moral significance, simply stated means that for an agent to be blameworthy they must be aware of the moral significance of their action; what it is about the action that gives it a moral dimension. If John intended to switch on the treadmill but was unaware, (due to, for example, upbringing or culture), that causing foreseen harm to Mary was wrong, then, (controversially), such ignorance could be claimed to eschew blame.

Third, the requirement of awareness of consequences. Continuing with the example of John, Mary and the treadmill; for John to be considered blameworthy he must hold a belief that at the time the switch was moved there was a reasonable possibility that the treadmill would move, and Mary would fall. There is controversy concerning how detailed and/or certain the belief must be to satisfy this requirement. For example, is the belief that Mary could-just-possibly fall sufficient for blameworthiness or is the stronger reasonable foreseeability needed for blameworthiness?

Fourth, the awareness of alternatives requirement. This requirement obviously touches much earlier discussion. Suffice at this point to say that an agent would be considered blameworthy, (for Marty's fall), if they believed that at least one alternative action was available and understood the consequences of that action, (that Mary would not fall).

[183] See Holly Smith's Paper *Culpable Ignorance* (1983) for very comprehensive consideration of this point.

The above introduces four epistemic requirements for moral responsibility in terms of the *contents* of awareness, but what *kind* of awareness is involved, what sort of mental states? The term 'epistemic' suggests that the kind of awareness involved should be described as knowledge. John would be clearly responsible and blameworthy if he *knew* the function of the switch and the consequences of switching. However, some philosophers argue that knowledge in the full sense is not required for blameworthiness, claiming that true belief, (rather than justified true belief) is required (Peels 2014: 493). For others, simply acting on the belief that the action is wrong is sufficient for blameworthiness (Levy 2011: 142). From this consideration the question arises, how are these beliefs, of whatever structure, to be characterised such that the agent has awareness appropriate for responsibility and blame or praise? There are two main options; either occurrently (Levy 2011: 141) or dispositionally (Levy 2014b: 34).[184] Essentially, the occurrent position reflects the initially plausible position that to be blamed or praised, (and satisfy the epistemic condition), for an action the agent must *at the time of the action* consciously believe that the action is wrong or right, and consciously consider some of the action's consequences. If these conditions are satisfied, then praise or blame is appropriate. However, consider the issue of culpable ignorance; John simply does not bother to understand the switch layout, even though he knows that understanding is an important safety requirement, leading to John's incorrect action and the harmful consequence of Mary's fall. If John *at the time of the action* consciously believed his action to be correct and consciously considered some of the action's consequences, then the action on the basic occurrent account would not be blameworthy. This is clearly implausible and so arises the idea of culpable ignorance and the notion of blameworthiness tracing back to previously blameworthy actions by the agent.

The disposition perspective concerning how beliefs are to be characterised essentially argues that the occurrentist position does not identify as blameworthy those agents who very plausibly deserve to be subject to reactive attitudes. This is particularly

[184] Following closely *The Stanford Encyclopedia of Philosophy* (Rudy-Hiller 2018). It is interesting to note, Gregg Caruso describes Neil Levy *Consciousness and Moral Responsibility* (2014b) as 'the most comprehensive and clear-headed examination of the relationship between consciousness and moral responsibility in the literature to date' and continues by providing a clear and concise summary of Levy's arguments in defence of the 'consciousness thesis', i.e., that 'consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility' (Caruso 2015).

the case if an occurrentist position is adopted whereby the requirement of (mistaken) belief that the action is a right action is not included; there is no consideration of rightness or wrongness at all. For example, in cases of forgetting that lead to harmful consequences there is no consideration of rightness or wrongness, or consequences by the agent *at the time* of the failure to act, hence on the occurrentist account there is implausibly no attribution of blame. These considerations lead to a revised characterisation of awareness within the dispositionalist account whereby 'tacit, dormant, dispositional or unconscious beliefs can … amount to the kind of awareness that is required for moral responsibility' (Rudy-Hiller 2018).

Concluding (i) describing the essential features of the debate concerning the epistemic condition, there are essential *agreements* among participants of the debate to be noted, described by Jan Willem Wieland (2017: 10) as 'the orthodoxy'. The focus here is culpable ignorance and following Holly Smith's (1983: 547) terminology the intention is to describe the relationship between the act or omission that leads to culpable factual or moral ignorance (the benighting act A1) and the later act carried out in culpable ignorance by an agent S (the unwitting act A2). It is claimed that 'all agree that an excuse by ignorance *can* render an agent blameless' (Wieland 2017: 10). However, when S is ignorant that A2 is wrong due to an absence of factual or moral information, then it is assumed by most participants that; (i) S is blameworthy for A2 only if S is blameworthy for being ignorant that A2 is wrong, and (ii) S is blameworthy for being ignorant that A2 is wrong iff S is blameworthy for a benighting act A1 that led to A2 (following closely Wieland 2017: 10). While described as the orthodoxy, there are several dissenting voices that argue, for example, that S is only responsible and blameworthy for the *benighting act* A1 that leads to A2, not A2 itself.

The above briefly describes aspects of the landscape within which discussion of the epistemic condition takes place. Discussion, for example, concerning how blameworthy overall is the culpably ignorant agent when the unwitting action does occur (A1 then A2) and when the unwitting action does not occur (A1 only), a situation that may be the result of luck alone.

(ii) Consider what John Martin Fischer says about the epistemic condition and guidance control, looking at 'particular facts surrounding an action' followed by, do they act with 'the proper sort of beliefs and intentions'.

I believe that Fischer and Ravizza's characterisation of the epistemic condition is not sufficiently detailed in the context of implicit bias and the questions raised by this Thesis. For example, Fischer and Ravizza's characterisation does not consider an important issue within the epistemic debate, one familiar and relevant during discussion of implicit bias; 'are ignorant agents morally responsible for their conduct?' (Wieland 2017: 2). Mention of insufficiency is clearly not a criticism of John Martin Fischer's account of guidance control, as previously noted, it is the control or freedom condition that is the focus of *Responsibility and Control* (1998), not the epistemic condition.

To examine the epistemic condition for guidance control in the context of implicit bias, I propose to employ a more comprehensive and detailed characterisation of the epistemic condition than agent awareness of 'particular facts surrounding an action' (Fischer 1998: 13). I believe it is reasonable to expand John Martin Fischer's basic epistemic requirement for guidance control, that an agent be aware of the 'particular facts surrounding an action' in terms of the four requirements given above; awareness of action, moral significance, consequences, and alternatives (Rudy-Hiller 2018). Similarly, I propose that John Martin Fischer's epistemic requirement for guidance control for agents to act with 'the proper sort of beliefs and intentions' be expanded and clarified using the ideas introduced above during discussion of the *kind* of awareness required by the epistemic condition. If it is found that implicit bias does not satisfy some aspect of the expanded epistemic conditions for guidance control, then clearly this will be looked at very carefully.

To summarise so far, it is the fourth aspect of guidance control, the epistemic condition, that is currently being considered. The essential features of the debate concerning the epistemic condition have been described and John Martin Fischer's position has been noted and a reasonable expansion proposed and described. I will now look at item (iii) using John Martin Fischer's *expanded* epistemic condition.

(iii) Consider if issuing behaviour of implicit bias satisfies John Martin Fischer's expanded epistemic condition, (hereafter 'expanded' will be assumed), for guidance control and responsibility.

First, I will examine the requirement that an agent be aware of the 'particular facts surrounding an action' now considered to be made up of four parts; an agent must be aware of (a) the action, (b) it's moral significance, (including comments on the moral sense and responsibility while unaware of the moral significance of reasons), (c) the consequences of the action, and (d) available alternative actions.

Second, attention will then turn to the epistemic requirement for guidance control that agents act with 'the proper sort of beliefs and intentions' using the ideas introduced above during discussion of the *kind* of awareness required by the epistemic condition.

To begin the first task, each of these aspects of John Martin Fischer's epistemic condition will be examined to confirm if implicit bias issuing behaviour is such that the epistemic condition is satisfied.

(a) An agent must be aware of the action. Awareness, responsibility and implicit bias have been discussed previously at some length. The discussion concluded that there are substantial and compelling arguments and empirical evidence that support the claim that agents *are* to a greater or lesser extent aware of implicit bias issuing behaviour. As this is such a fundamental point I will reinforce it with reference to Bertram Gawronski *Six Lessons for a Cogent Science of Implicit Bias and Its Criticism* (2019). Gawronski's Paper responds to increased scrutiny, sometimes scepticism, of the explanatory value of the implicit bias model, suggesting six 'lessons' for an informed science and critical debate of implicit bias. The first lesson discusses the claim that 'there is no evidence that people are unaware of the mental contents underlying their implicit biases' (2019: 575).[185] There

[185] Gawronski develops an important point at the outset of this section; implicit measures, in contrast to explicit measures, do not *require* that participants are aware of the to-be-measured mental contents, however, 'it is often assumed, on the basis of this *methodological difference*, that explicit measures capture conscious biases, whereas implicit measures capture unconscious biases'. Gawronski continues, 'Because implicit measures do not *require* awareness of the to-be-measured mental contents, they certainly have the potential to capture unconscious mental contents that evade assessment via explicit measures. However, this possibility *does not* imply that people are unaware of the mental contents underlying their responses on implicit measures. Any such claim is an empirical hypothesis that has to be evaluated on the basis of relevant evidence. Indeed, a closer look at the *available* evidence raises serious doubts about the veracity of this hypothesis' (Gawronski 2019: 575).

are three conclusions: (i) Statements about unawareness should be treated as hypotheses that require empirical evidence. Implicit biases have multiple aspects that could be outside of awareness, therefore it is essential to clearly specify which aspect is assumed to be outside of awareness.[186] (ii) Counter to a widespread assumption in the literature, there is currently *no evidence* that people are unaware of the mental contents underlying their responses on implicit measures. The available evidence suggests that people *are aware* of the mental contents underlying implicit measures, which allows them to predict their implicit-bias scores with a high degree of accuracy.[187] (iii) The same conclusion applies to claims about lack of source awareness *and* lack of *impact* awareness, which should be assessed with appropriate designs and reliable measures of awareness. At this point, the available evidence suggests that people can be unaware of the *origin* of their implicit biases, but the same is true of explicit biases. Moreover, the preliminary evidence that implicit, but not explicit, biases influence judgments and behaviour *outside* of awareness is rather weak and prone to alternative interpretations. (2019: 578)

Based on Gawronski's conclusions (i), (ii) and (iii) above and earlier discussion, is it reasonable to claim that the epistemic condition of *awareness* of action is satisfied when an agent acts as a result of implicit bias? It is surely reasonable to continue, as a minimum, with the assumption of content and impact awareness by the agent. Following conclusion (iii), source awareness, the origin of the underlying mental contents, will be left as an open issue for the moment.

(b) An agent must be aware of an action's *moral significance*.

Confirming the way forward, in this section I will look at item (b), followed by some comments on the moral sense and its response to immoral practice within an endorsing

[186] Recall from page 167 (a) Content awareness; the mental contents underlying responses on implicit-bias measures, (b) Source awareness; the origin of the underlying mental contents, or (c) Impact awareness; effects of the underlying mental contents on judgments and behaviour.

[187] Please note: The essential arguments and conclusion of this Thesis, (i.e., from the semicompatibilist perspective, agents *are* responsible for their behavioural expression of implicit biases), do not depend on the truth or falsity of agent awareness of the mental contents underlying implicit measures. This is supported, for example, by Deutsch and Strack's confirmation that 'the partition of the reflective–impulsive model (as adopted within this Thesis) into two systems is *not* based on the presence *or* absence of conscious awareness' (added addition and emphasis Strack and Deutsch 2004: 238).

culture based on *Blame and Moral Ignorance* (Sher 2017).[188] Then, I will comment on the question, is it possible to be considered responsible for an action and be *unaware* of its moral significance? These issues cannot be explored in depth here but are mentioned because they connect with aspects of implicit bias and responsibility: The possibility of the moral sense[189] as a response to instances of implicit bias influenced behaviour within an endorsing culture, and responsibility for an action when unaware of its moral significance clearly links with the epistemic condition to be blameworthy an agent must *be* aware of the moral significance of their action. After completing, I will return to the remaining epistemic conditions; (c) the consequences of the action, and (d) available alternative actions. When all of these items are complete it will be possible to briefly summarise findings concerning the condition 'an agent is responsible only if he knows the particular facts surrounding his action. Attention will then turn to the second main task, examining the epistemic requirement for guidance control that agents act with 'the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13) using the ideas introduced above during discussion of the *kind* of awareness required by the epistemic condition. After making further comments on implicit bias and degrees of awareness and responsibility it will be possible to make an overall assessment of *all* the requirements of guidance control in light of implicit bias. Finally, I will explore moral luck and semicompatibilism, an issue touched upon during discussion of the formation of an agent's operating mechanism and moral sense.

So, to continue, item (b), moral reasons, as a necessary part of guidance control, has been mentioned during discussion of John Martin Fischer's second condition for guidance control (2.a.ii.).[190] Commenting on this condition I said that implicit bias is

[188] See also Elinor Mason and Alan T. Wilson *Vice, Blameworthiness, and Cultural Ignorance* (2017). This Chapter, from Robichaud and Wieland ed. book *Responsibility: The Epistemic Condition*, argues ignorance *can* be culpable even in situations of widespread cultural ignorance.

[189] See Cudworth (1996) and Hutcheson (2017) for detailed discussion of the moral sense. Also, Paul Russell's vital Paper, Moral Sense and the Foundations of Responsibility *The Oxford Handbook of Free Will: Second Edition* (2011: 199-220); a critical assessment of Strawson's project and the role of the moral sentiments or reactive attitudes (2011: 200).

[190] Item (2.a) of John Martin Fischer's requirements for guidance control states: K is regularly receptive to reasons, some of which are moral. This requires; (i) That holding fixed the operation of a K-type mechanism, the agent would recognize reasons in such a way as to give rise to an understandable pattern

clearly charged with morally related content, including morally relevant qualities such as untrustworthiness, cowardice and so on (Eberhardt 2019: 31). This comment still stands, but there is a further point to make; being *receptive* to reasons some of which are moral (2.a.ii) and being aware of the moral *significance of* such reasons, having 'a belief about the presence of whatever features *make the action wrong*' (Rudy-Hiller 2018), are not the same thing. The epistemic condition of being aware of the moral *significance of* such reasons requires *understanding* by the agent of what makes particular actions right or wrong. At the higher levels of the chosen model of implicit bias there is awareness and reflection (see Fig 5.1 and related discussion), with greater or lesser agent responsibility for issuing actions. However, is the *level* of *moral* awareness experienced by the agent when actions are influenced by implicit bias sufficient to meet the requirements of guidance control? Guidance control requires the agent to be receptive to reasons, some of which are moral (2.a.ii), but the epistemic condition (item 4) in its expanded form calls for more substantial awareness of the moral *significance* of such reasons. How is this to be reconciled? A further question is also suggested, is it possible to be considered responsible for an action and be completely *unaware* of its moral significance? Exploring these questions will be the next task and, as mentioned above, will include consideration of the 'moral sense'.

Implicit bias is often described in terms such as operating beneath our awareness (Eberhardt 2019: 170), (Staats, Cheryl and others 2017: 10) but a model of implicit bias has been presented having varying degrees of awareness and reflection by the agent of the content and impact of behaviour. Given a level of awareness, implicit bias related actions, like other actions, take place within the context of the agent's background moral beliefs. Typically, such beliefs include 'the moral importance of truth-telling and promise-keeping, treating others fairly, not harming, and providing aid when it can be done at acceptable cost and risk' (Sher 2017: 112). It is reasonable then, to claim the level of moral awareness experienced by an agent when actions are influenced by implicit bias varies according to the character and sophistication of the agent, (and the degree of stress and environmental factors affecting the agent). Given there *is* a degree of moral

from the viewpoint of a third party who understands the agent's values and beliefs, and (ii) That some of the reasons mentioned in (2.a) are moral reasons.

awareness relating to behaviour, (depending on character and so on), I do not see a distinction here between implicit bias influenced behaviour and behaviour originating from other sources. Take, for example, the general claim that if a morally ignorant wrongdoer satisfies all non-epistemic conditions for blameworthiness, he is blameworthy for acting wrongly if, but only if, he was at least in a position to recognize that his act was wrong when he performed it (Sher 2017: 102); such a claim is equally appropriate for implicit bias related behaviour and behaviour originating from other sources. When a level of content and impact awareness is admitted for implicit bias, then moral awareness is surely common across sources of behaviour for everyone. Consider Levy's point when he says, 'In order for an agent to be morally responsible for an action, it is not sufficient that the mechanism causing his action is sensitive to a broad range of reasons; it had better be sufficiently sensitive to the kind of reasons that give to his actions their moral character' (2017: 17). Sensitivity to the kind of reasons that give (to his) actions their moral character is present within the model of implicit bias for behaviour influenced by implicit bias and other sources.

The 'depth' of moral awareness varies according to the character and sophistication of the agent and the degree of stress and environmental factors affecting the agent. But understanding the moral *significance* of reasons (or actions) is not all or nothing and variation in understanding is perhaps reflected (along with other considerations) in the degree of agent responsibility. While some agents may have the capacity for deep reflection on the moral *significance of* reasons, I believe John Martin Fischer's relatively 'simple' statement, that a responsible agent must be (at least) receptive to reasons, some of which are moral reasons, to be sufficient; there *is* a moral dimension or moral consideration within the reasons responsive mechanism.[191] Some agents may have the capacity to reflect on the deeper significance of moral reasons, but as stated, I believe to be receptive to moral reasons is sufficient to act with guidance control and so responsibility. Further, the 'receptive to reasons, some of which are moral

[191] It is tempting to try to frame differences between awareness of moral significance and being receptive to reasons in terms of a *de re / de dicto* distinction. In general terms, *de re*: a desire to act on actual, correct, moral principles and to do what is in fact the right thing, and *de dicto:* a concern for doing what one *feels or believes* that one morally ought to do. See George Sher, *Blame and Moral Ignorance* (Sher 2017: 103) for discussion of Quality of Will theories and a thorough description of *de re / de dicto.*

reasons' requirement could be framed in terms of agents' moral sense in situations where, for example, implicit bias is influential. Of course, there are issues concerning the *formation* of the agent's moral sense, how the sensitivity for what is right or wrong or what one *feels or believes* one morally ought to do has been formed, and continues to be formed, over time. More specifically, formation of the moral sense is surely not immune from the very cultural influences that probably contribute to the formation of implicit biases? Considering past cultures that embraced a morality allowing, (perhaps endorsing), behaviour that today is considered wrong, is it reasonable or possible that an agent's moral sense should or could have responded? This is an important point, and I will briefly make some comments on the moral sense in the context of immoral practice within an endorsing culture. The important general problem of luck, particularly constitutional luck and the threat to semicompatibilism/guidance control will also be discussed later in this chapter.

I will continue with some ideas about the moral sense and immoral practice within an endorsing culture, where behaviour was based on values that today are considered wholly in error. How reasonable is it to expect those living within such cultures to recognise moral errors, (relative to absolute or our contemporary cultural values), and do something about it? George Sher (2017: 113) makes the point that given an agent *is* reflecting on some form of activity within their society, any thoughts that such activity could be wrong may be opposed by pervasive counter opinion. It would be unreasonable for the agent to ignore the opinions of members of their society, particularly if such opinions are in the majority, held by many who are both intelligent and widely respected and/or supported by ideologies or religious doctrines (following Sher 2017: 113). Sher sees the reflective process in terms of weighting; the balancing of reasons, some of which are prevailing attitudes. Sher continues interestingly as follows; Over time the weight an agent attaches to the significant opinion of others within their society declines. Importantly, Sher suggests there are certain beliefs and values *that everyone holds*, for example, that it is wrong to deprive others of their liberty or lives, even in societies that allow or require slavery.[192] This is certainly a controversial claim; it just seems to be the

[192] See also the Journal article *Historical Roots of Implicit Bias in Slavery* (Payne, Vuletich and Brown-Iannuzzi 2019). This article explores an interpretation of implicit bias as the cognitive residue of past and present structural inequalities. Specifically, the article investigates the historical persistence of implicit bias,

case that some people *truly* believe, for example, that certain races have natural ascendancy and have sought to prove their beliefs to be objectively true using pseudo-scientific methods (Eberhardt 2019: 133). However, given there *are* values that everyone holds, it is suggested the agent will *always* question, to a greater or lesser extent, why the majority do not respond to these common values and will assume there must be another, more important, reason that has greater weight that supports and justifies the actions and beliefs of others apart from their pervasive nature. As time passes no such reason becomes visible and the agent assigns greater weight to their *own* sense of wrong. A tipping point is reached where the agent's own sense of wrongness is in the ascendancy; the balance of reasons becomes such that objectionable aspects of the agent's society are rejected. Using a Jo Jo type example (Wolf 2003: 379), someone who from childhood has been indoctrinated such that they hold the worst kind of beliefs and values, Sher asks whether such a person could *ever* have access to reasons for rejecting or even questioning their beliefs, and concludes the answer is yes. This conclusion is based on the claim that even in extreme cases people have access to the requirements of 'common-sense morality' (Sher 2017: 114), (although the term 'common' can seem questionable in this context). Importantly, Sher notes the distinction between reflecting on the wrongness of various actions and actually progressing such reflection into *action* due to '… social pressure, personal advantage, intellectual laziness, fear, and simple inertia …' (Sher 2017: 114). An agent on Sher's account does have *access to reasons* for rejecting or questioning their beliefs; for Sher, the epistemic condition of the agent is such that they *are* culpable for their wrong behaviour even within a culture that supports such wrongdoing. The issue of culpability here seems to rest on whether environmental conditions can *eliminate* the possibility of an agent accessing fundamental values and reasons that could lead towards better behaviour. Do such 'fundamental values' exist in a real sense and why should they have a particular nature and be shared across time and cultures? Questions that cannot be considered here but are nonetheless clearly and fundamentally important. Further, is it true that reflection and reason (necessarily) lead to the conclusion that certain actions and behaviour are wrong? Intriguingly, Rebecca

comparing levels of current implicit bias with the proportion of the population enslaved within states of the USA in 1860, (summary based on article Abstract).

Goldstein Newberger asks why Plato never questioned slavery, claiming 'Plato, no more than any other ancient Greek, including his brilliant student Aristotle, never thought to question the institution of slavery, the whole abominable notion of one person owning another' (2015). If true, this suggests the beginning of an interesting and persuasive counter to Sher's argument, that reason and reflection ultimately reveal mistakes in moral judgement. From Sher, it seems reasonable to expect at least a far stronger reaction from Plato against slavery than is apparent within his Works. While the description 'never thought to question the institution of slavery' is a simplification, Benjamin Jowett does comment in his introduction to *Laws* that for Plato, slaves

> are to be treated with perfect justice; but, for their own sake, to be kept at a distance. The motive is not so much humanity to the slave, of which there are hardly any traces […] but the self-respect which the freeman and citizen owes to *himself* […] Plato still breathes the spirit of the old Hellenic world, in which slavery was a necessity, because leisure must be provided for the citizen.[193] (added emphasis Plato 2019: 10754)

There is much that could be said concerning Plato's position on slavery and the role of reason (and emotions) in moral progress, but *as a minimum* and contra Sher, the example of Plato suggests that it is not clearly the case that agents are culpable for their behaviour within a culture that supports (or is indifferent to) such wrongdoing, when such a claim is based on the availability of universal values by application of reason. When reasoning correctly, is it inevitable that true and real values emerge that challenge existing practices or is it the case that agents can reason perfectly well but based on bad influences and/or incorrect information reach conclusions that with the benefit of hindsight are grossly wrong relative to current thinking, (of particular societies, and to a greater or lesser extent). What can be concluded concerning independence of the moral sense from cultural influences? Such brief consideration here can only offer, at best, sketchy comments on what is, I believe, essentially a discussion of the substantial issue of moral realism and moral anti-realism and the extent that values are accessible to reason and

[193] 10754 is the location reference of the quotation from the Kindle Edition of The Essential Plato Anthology (25 works), translated by Benjamin Jowett.

reflective enquiry. A realist view,[194] that moral rightness or wrongness is, in an important sense, a claim about matters of fact which are either true or false, has great intuitive appeal. For example, the proposition slavery is wrong is just true, it is surely a factual claim. However, the question remains, by reflection and reason, is it inevitable or even likely that true and real and/or objective values emerge that challenge existing practices?

From the perspective of Kant, that knowledge of moral law *is* accessible to every rational person by virtue of their rationality and applies to all rational beings in any world, is a striking claim considering the earlier question; do (such) 'fundamental values' exist perhaps in some real sense and why should they have a particular nature and be shared across time and cultures? Kant's well-known notion of a morally permissible action necessarily based on a maxim that could be applied as a universal law, for example, that human beings should be treated as an end in themselves and not as a *means* to an end, is again striking in the context of the earlier discussion of slavery. Kantian ethics clearly offers a very attractive departure point for more detailed consideration of these issues.

I believe the requirement of guidance control, being 'receptive to reasons, some of which are moral reasons' is satisfied when agents exercise their moral sense in situations where implicit bias could be influential. The issues concerning the *formation* of an agent's moral sense, *how* concern for doing what one feels or believes one morally ought to do is formed and unease over moral relativism remain unsettled and controversial.[195] The formation of an agent's actual mechanism over time is subject to similar concerns, being influenced, for example, by the culture the agent happens to be immersed within. These matters are considered later from the perspective of moral luck. Given that the constitution of an agent's mechanism is influenced to some extent by lucky circumstances, is there sufficient control to warrant responsibility?

[194] I have used the term 'realist' but recognise the debate concerning being realist about X, where one believes X is objective, and believing X is objective and yet not being realist about it. If by realist/realism one means something like mind independent or independent of any conceptual scheme or frame of reference; one might think that the rules of chess or tennis are objective, while not being real in the mind independent sense. In light of this, it would be more accurate to say I find an objective view very plausible. (Following *The Electric Agora,* Value and Objectivity by Daniel Kaufman (2020).

[195] See for example Ralph Cudworth *A Treatise Concerning Eternal and Immutable Morality* (1996) for detailed discussion of the principles of morality, knowledge and metaphysical realism. For an alternative perspective see Francis Hutcheson *An Inquiry Into The Original Of Our Ideas of Beauty And Virtue* (2017).

Recall the outline of the way forward presented earlier (page 176); the next item to consider is possible responsibility for an action when an agent is *unaware* of its moral significance.

Earlier brief discussion of the moral sense and immoral practice within an endorsing culture concerned an agent *unaware* of (or desensitised to) the moral significance of an action due to the concealing effect of pervasive views within their society that normalise and characterise morally wrong actions or behaviour as natural or part of a divine order.[196] Is it possible that an agent *could* fail to recognise the moral significance of certain behaviour; if it is possible, then what sort of arguments, if any, could justify blame when an agent is *just not aware* that particular behaviour has a moral perspective? The idea that the moral perspective of particular behaviour can be concealed from an agent by pervasive cultural and environmental influences is essentially rejected by Sher (2017). On this view, universal values *cannot* be entirely suppressed or masked by culture. (That there is *variation* in degree of bias across a society and the extent that bias is manifest suggests other factors are in play, factors that *could* be within agent control such as explicit beliefs (following Holroyd 2012: 281)). More generally, values cannot be entirely suppressed or masked by *any* condition, therefore not being aware of a moral perspective is impossible. In the final paragraph, Sher's position changes, claiming that 'each of us is sure to have many moral blind spots that he is simply not in a position to identify. […] we can all predict with confidence that we are likely to be blameworthy for many acts that we do not currently consider wrong' (2017: 116). If there are 'moral blind spots' then surely access to at least some (universal) values is denied, but previously Sher argues that agents were blameworthy at the time of wrongdoing because such fundamental values *are* ultimately available to the agent and so potentially challenging their behaviour. Is it possible to be responsible for an action *and* be unaware of its *moral* significance? Being blameworthy for an action when *culpably unaware* of its moral significance is plausible. Being blameworthy for an action when simply *unaware* of its moral significance is implausible. If it is the case that moral rightness or wrongness is a claim about universal matters of fact, is an agent's failure to recognise

[196] See Jennifer Eberhardt (2019: 133) for discussion of the role of nineteenth century science in promoting the view that inferiority of certain races, and so their treatment, was part of the natural (created) order, confirmed by false scientific methods.

(and act on) such facts culpable? If it is, (following Sher's main argument) an agent would be considered blameworthy for related behaviour. However, if an agent has a 'blind spot' for a particular moral value, then being simply *unaware* of the moral significance of the related action surely leads to behaviour that is not blameworthy. So, is it possible to be responsible for an action *and* be unaware of its *moral* significance? Yes, if unawareness is culpable because of, for example, laziness or self-interest. Also, yes, if an agent has through an earlier benighting act failed to seek out relevant challenging universal values available through reflection.

I do not believe it is necessary to reach a conclusion on these many points to respond to John Martin Fischer's epistemic condition, that (to be subject to guidance control) an agent must be *aware* of an action's moral significance. John Martin Fischer does not talk about an action's moral significance in relation to universal *or* contemporary values, simply an awareness, (of an action's moral significance). John Martin Fischer's second epistemic condition does not take a position on the moral realism – relativism debate, so this substantial, important and difficult problem can be respectfully moved to one side; John Martin Fischer only requires that an agent must be *aware* of an action's moral significance in the context of their society's value conventions *or* perhaps some objective standard. When reflecting on a possible action an agent *will* surely be at least aware, and would probably consider, the action and associated values, norms and conventions of their society. While an agent may choose to ignore such conventions in their decision making, it seems implausible that they would be unaware of their own society's essential value conventions. In the unlikely event that an agent *was* unaware, (of an action's moral significance), then following John Martin Fischer's condition, they would not be responsible, unless unawareness was culpable.

Concluding section (b), that an agent must be aware of an action's moral significance, clearly as noted, implicit bias is charged with morally related content, including morally relevant qualities, but is the level of moral awareness experienced by the agent when actions are influenced by implicit bias sufficient to meet the requirements of guidance control? The requirements of guidance control were found to be unclear in terms of 'receptive to reasons some of which are moral' or the more substantial awareness of the 'moral significance of reasons'. It was concluded that while some agents may have the capacity for deeper reflection on the moral significance of reasons, (perhaps suggesting a greater degree of responsibility), John Martin Fischer's relatively 'simple' statement, that a responsible agent must be (at least) *receptive* to reasons, some of which are moral reasons, is at least sufficient as a foundation of agent responsibility. Is the level of moral awareness experienced by the agent when actions are influenced by implicit bias sufficient to meet the requirements of guidance control? The adopted model of implicit bias includes the possibility of reflection, awareness and so responsibility. Moral awareness, and being receptive to moral reasons, is naturally integrated with this model of mechanisms that mediate the influence of implicit social cognition on behaviour (fig 5.1). The moral sense, and responsibility in the absence of awareness of moral significance were discussed, as they are both issues connected with implicit bias and responsibility. In the unlikely event that an agent *is truly* unaware, not having even a flicker of comprehension within their construal of a situation (of an action's moral significance), then following John Martin Fischer's condition, they are not responsible, (unless unawareness was culpable).

I will now turn to John Martin Fischer's other epistemic conditions; that an agent must be aware of; (c) the consequences of the action, and lastly (d) available alternative actions.

(c) An agent must be aware of the consequences of the action. It seems correct to say that to be responsible for the outcome of an action an agent *must* have held *at the time* a belief about the *possibility* of that outcome occurring as a consequence of that action, (or be culpable in some sense for not holding that belief). As referred to earlier, there is some dispute concerning how detailed the belief must be. In the earlier example, to be responsible, does John at the time of the action have to believe it *reasonably foreseeable* that Mary may be hurt, or to be responsible is a more specific belief required, such as a belief that her arm may be broken?[197] When discussing tracing, Fischer and Tognazzini express the point, ' … there will be a range of specifications, each more coarse-grained than the previous, and while some will not have been reasonably foreseeable, others will (2009: 537).

[197] See Fischer and Tognazzini *The Truth about Tracing* (2009) for similar discussion.

How is a belief about the possibility of an outcome with implicit bias as its source to be understood? From everything that has been discussed concerning awareness, implicit bias and issuing behaviour, the idea of having belief about the possibility of outcomes with implicit bias as the source seems reasonable and plausible. Given a reflective element within the model of issuing behaviour, it is difficult to understand how belief about possible outcomes could fail to be a part of such reflection. The responsibility endorsing epistemic condition of awareness of the consequences of an action, where an agent must hold at the time a belief about reasonably foreseeable outcomes of an implicit bias issuing action, I believe is plausibly satisfied. Even limited reflection on actions must *surely* include some beliefs about possible future outcomes. I will now turn to the fourth and final condition, item (d).

(d) An agent must be *aware* of available alternative actions. Initially, this seems a reasonable condition; unaware of alternative actions an agent surely cannot be blamed for taking what is believed to be the only course available, (assuming taking no action is impossible)? However, there is more to consider before leaving this item. This condition evokes earlier discussion of responsibility, determinism and the whole idea of responsible action if it happens to be the case that alternative paths are not available. In a previous example, an agent believed alternative actions were available, when in fact there were no alternatives; unaware that the door of their room had been locked, they just happened to choose to stay in the room. Crucially in this case the agent was responsible for the act of staying in their room as this action took place via *the agent's own* responsibility relevant mechanism. The state of being (in fact) unable to do otherwise did not mitigate responsibility; the agent believed (assumed) alternative actions were available but chose not to take them and fully endorse the action that *was* chosen. If an agent believes *no* alternative action is available, (whether a true belief or not), yet endorses that action, making it their own, then the agent is responsible for that action. If an agent believes that no alternative action is available, (whether a true belief or not), and does *not* endorse that action, feeling coerced or forced in some way to act, (assuming taking no action is not an option), then responsibility for that action (other than in a minimal sense) becomes *very* questionable.[198] John Martin Fischer supports this claim, listing 'force' as the 'second type of excusing condition' (1993: 7). Iff an agent *owns* the action via *their own* responsibility relevant mechanism does the agent become responsible *for* that action.

So, an agent is responsible for their actions if *aware* of available alternatives they make and endorse a choice, making it their own, *and* if not aware of alternative actions, an agent takes the only action available while endorsing it as their own. This clearly suggests a point to be made concerning the above requirement for a responsible agent to 'believe in or be aware of available alternative actions'. If an agent owns the action's delivery mechanism, (and has 'put to the side' any doubts about their responsibility establishing agency (Fischer and Ravizza 1998: 62-91, 243-244)), then awareness of or belief in possible alternative actions is not required.

The expanded version of John Martin Fischer's epistemic condition is plausible, and this additional point does not undermine or effect any earlier conclusions or discussion and does not cause a problem. It happens to be the case that John Martin Fischer's guidance *control* condition of *owning* and taking responsibility for the issuing mechanism confers responsibility whether alternative actions are actually available or believed to be available. The control condition (owning and taking responsibility for the issuing mechanism) is the essential requirement; John Martin Fischer makes it clear that actual or believed access to available alternative actions is an unnecessary condition for guidance control, it may be 'put to the side'. There is no intention to undervalue the sense that from the inside 'alternative possibilities seem to lie open before us … and one of the possibilities is made actual by what we do' (Nagel 2003: 232). This is the reality of our *experience* of the world, and as Nagel describes, part of the problem when trying to understand autonomy and responsibility.

The situation may be summarised in quite a straightforward way; the adopted model of implicit bias facilitates agents' varying degrees of awareness and reflection concerning (actual or illusionary) alternative actions and choices, our sense that from the inside alternative possibilities appear to lie open before us. However, for guidance

[198] An agent's *beliefs* about the implications of the truth of determinism (availability of alternatives) is discussed by John Martin Fischer *Responsibility and Control* where it is suggested that it is 'plausible that individuals can be brought to take a certain sort of stance in which metaphysical doubts are put to the side (practically speaking)' (1998: 229).

control such awareness of alternative actions and choices is *un*necessary if an agent owns the action's delivery mechanism and has 'put to the side' any doubts about their responsibility establishing agency even in light of, for example, the possible truth of determinism.

Briefly summing up so far, for the agent to be aware of 'particular facts surrounding an action' the agent must be aware of action, moral significance, consequences and alternatives' (Rudy-Hiller 2018): (a) The action; there are substantial and compelling arguments and empirical evidence supporting the claim that agents *are* to a greater or lesser extent aware of implicit bias issuing behaviour i.e., the agent is aware of the action. (b) The action's moral significance; moral awareness, and being receptive to moral reasons, are naturally integrated within the adopted model of mechanisms that mediate the influence of implicit social cognition on behaviour and so available as a component of guidance control. (c) The consequences of the action; the responsibility endorsing epistemic condition of awareness of the consequences of an action, where an agent must hold at the time a belief about reasonably foreseeable outcomes of an implicit bias issuing action, I believe is plausibly satisfied. (d) Available alternative actions; it has been argued that implicit bias issuing actions *are* subject to varying degrees of reflection and accompanying awareness of (actual or illusionary) alternative actions and choices. It was noted that for guidance control such awareness of alternative actions and choices is *un*necessary if an agent owns the action's delivery mechanism and has 'put to the side' any doubts about their responsibility establishing agency (for example, in light of the possible truth of determinism). From Nagel it was noted that from the 'inside' there *is* an important sense of alternative actions.

Next, I will look at John Martin Fischer's second general requirement, that an agent 'acts with the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13); does an agent act with 'the proper sort of beliefs and intentions' when actions are directly influenced by implicit bias?

What do Fischer and Ravizza mean by 'the proper sort of beliefs and intentions'? The context of this requirement is *Responsibility and Control,* Chapter Eight, Section II, Becoming a Moral Agent:

> Parental responses to a child's behavior, as part of the typical process of moral education, seek to induce the child to accept a certain view of himself as an agent. The relevant notion of 'agency' is a rather minimal notion, according to which the child sees himself as the source - in a specific sense - of certain upshots in the external world. The sense in which the child sees himself as the 'source' of these upshots is that he sees that their occurrence is caused - in a certain characteristic way - by him. The child is brought to see that his desires, *beliefs, and intentions* result in actions and upshots in the world; these upshots are not the results of freakish accidents or other agents. (added emphasis Fischer and Ravizza 1998: 208)

Fischer and Ravizza's position is that 'moral responsibility is an essentially *historical* notion; someone's being morally responsible requires that the past be a certain way' (1998: 207). Recall the ideas introduced above during discussion of the *kind* of awareness required by the epistemic condition, particularly the characterisation of awareness within the dispositionalist account whereby 'tacit, dormant, dispositional or unconscious beliefs can … amount to the kind of awareness that is required for moral responsibility' (Rudy-Hiller 2018). This runs with John Martin Fischer's account that the past must include a process of 'taking responsibility', a necessary feature of moral responsibility. 'It (taking responsibility) is part of the process by which a mechanism leading (say) to an action, becomes one's own' (1998: 207).[199] Recall that taking responsibility involves three major ingredients: First, seeing oneself as the source of one's behaviour in the quite minimal sense that one sees that one's 'desires, *beliefs, and intentions* result in actions and upshots in the world ... ' (Fischer and Ravizza 1998: 208). Second, one must see oneself 'as a fair target of the reactive attitudes as a result of how (one) exercises this agency (1998: 211). Third, these views of oneself must be based on evidence (1998: 213). By 'the proper sort of beliefs and intentions' Fischer and Ravizza mean, based on appropriate evidence, self-conception of agency that 'involves the minimal idea that the child is an agent *qua* source of certain upshots in the external world, such that his own desires, *beliefs, and intentions*, not freakish accidents or other agents, result in actions and upshots in the world.[200] The

[199] Questions have been raised concerning Fischer and Ravizza's claim that moral responsibility is an essentially *historical* notion. See *Review: Fischer and Ravizza on Moral Responsibility and History* (Bratman 2000).

[200] Description of 'the proper sort of beliefs and intentions' is in harmony with Thomas Nagel's Paper *Freedom* (2003); from the internal perspective, the sense of being the source of change in the world, (yet from the external perspective, the agent is part of the world 'and our lives are seen as products and manifestations of the world as a whole' (2003: 232).

reactive attitudes self-conception involves the child's belief that he is sometimes the appropriate object of the reactive attitudes' (Zimmerman 2002: 219). Initial reaction to this final requirement of guidance control, that an agent 'acts with the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13), is that it is both reasonable and straight forward. The minimal sense that our chosen actions bring changes in the world that may legitimately result in praise or blame from others is for most human beings an accepted and assumed situation. As discussed at length, from the perspective of implicit bias, there is controversy about whether the relevant desires, beliefs and intentions are *his own* desires, beliefs and intentions, (such questions may also be asked of desires and beliefs concerning many issues, from mundane to life changing). As previously concluded, (page 142), the desires and beliefs that issue in implicit bias related behaviour *do* belong to the agent in a responsibility enabling way: 'If agents exercise this form of control, (ecological control) … then implicit attitudes can be properly regarded as part of 'who the agent is' - part of her character, which is as a whole subject to moral evaluation' (Holroyd and Kelly 2016: 130). So, it is correct to claim that an agent's sense of being *qua* source of certain upshots in the external world that issue from their own desires, beliefs, and intentions, can apply to both implicit bias related actions and actions that have other sources. The semicompatibilist/guidance control requirement that an agent 'acts with the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13) may be satisfied in the case of implicit bias influenced behaviour.

Responding to the general question raised at the beginning of this section on the epistemic condition for guidance control; does issuing behaviour of implicit bias satisfies John Martin Fischer's epistemic condition of guidance control and responsibility? After detailed examination of an expanded epistemic requirement for guidance control in the context of implicit bias influenced behaviour it was concluded that the epistemic condition is satisfied.

Before drawing together and presenting a summary of the conclusions of the investigation into implicit bias and *all* the requirements of guidance control (items 1 through to 4), I will make some final comments on implicit bias and degrees of awareness and responsibility.

The argument that agents are excused blame for implicit bias related behaviour because implicit biases are unconscious or the result of irresistible subtle, (or less than subtle), cultural influences beyond the agent's control has been discussed, including issues such as awareness and so possible, (but not guaranteed) control of issuing behaviour. In response, and generally contra[201] to Levy (2014a), it is argued, for example by Holroyd, that 'there is some reason to suppose that individuals *are,* at least sometimes, liable to blame for the extent to which they are influenced in behaviour and judgment by implicit biases' (2012: 274). The implicit bias model described and used within this Thesis includes Holroyd's general thesis and accepts a particular model of the influence of implicit social cognition on behaviour (Fig 5.1) acknowledging the possibility of reflective behaviour. However, as Alex Madva points out, 'awareness and self-knowledge are often quite *difficult* and must draw on a rich store of observations and interpretations of … behaviour' (added emphasis 2020: 390). Difficulties experienced when making reflective behavioural decisions, perhaps due to stress, tiredness, or general cognitive load, leads to consideration of moral responsibility when there is *partial* awareness of implicit bias. If awareness comes in degrees, then can (plausibly and intuitively) control, responsibility and blameworthiness be meaningfully said to 'come in degrees' (Madva 2020: 390)?[202] It is this notion of *degrees* of awareness and responsibility that will be touched upon in this next section.

There are several reasons for mentioning degrees of awareness and responsibility. First, such consideration offers a way of capturing the idea that, for example, a racist living during the 18th Century, ' … while still morally responsible and blameworthy for their racist attitudes, is nevertheless less responsible and blameworthy for their attitudes than a contemporary racist … ' (Coates and Swenson 2013: 643). Second, the idea of degrees of awareness and responsibility may be expressed in terms of guidance control because the 18th Century racist is less reasons-receptive, less aware of liberal reasons for

[201] This is clearly a gross simplification of the relative positions and given merely as an introduction to discussion of partial responsibility and blame.

[202] See also Madva *Implicit Bias, Moods, and Moral Responsibility* (2018) where the implications of empirical evidence that individuals *are* aware of their implicit biases in partial and inarticulate ways is explored. It is argued (by analogy with moods) that responsibility and awareness, (and control), come in degrees and partial awareness of implicit biases makes (agents) partially morally responsible for them.

treating persons of all races as equal (Coates and Swenson 2013: 643). Coates and Swenson's Paper *Reasons-Responsiveness and Degrees of Responsibility* (2013) thoroughly develops the notion of degrees of responsibility expressed within the paradigm of guidance control. The authors outline the point as follows:

> …the greater degree of comparative similarity that obtains between the actual world and the nearest possible world in which the actual sequence mechanism reacts to sufficient reason to do otherwise, the greater degree of responsibility. And correspondingly, the greater the degree of blameworthiness. Likewise, if there is less comparative similarity between the actual world and the nearest possible world in which the actual sequence mechanism reacts to sufficient reason to do otherwise, the agent in question is less responsible. And correspondingly, agents whose actions issue from such mechanism will be less blameworthy for their actions. (Coates and Swenson 2013: 636)

Examples are given to clarify the point, but the above quotation shows the general argument and the intuitive idea that degrees of responsibility may be expressed within the semicompatibilist/guidance control paradigm. There is much that could be considered in Coates and Swenson's Paper in terms of guidance control and degrees of responsibility that is unfortunately beyond the main objective of this Thesis. See also Madva (2018: 63) for contra view of Fischer and Ravizza's position, that responsibility is binary even when awareness, control, blame and reason responsiveness all come in degrees.

The third and final point concerning degrees of awareness and responsibility returns to implicit bias, Deutsch and Strack (2010), motivation, opportunity and implicit and explicit measures; '… motivation and opportunity to engage in more complex reasoning processes have proven to moderate the relation between *implicit* and *explicit* social cognition on one hand and behaviour on the other' (2010: 66). Motivation or opportunity have a decisive influence on whether processes associated with implicit or explicit measures have the upper hand. Concerning *implicit* measures, the notion of degrees of responsibility is reflected within the model of implicit bias (Fig 5.1), where important mechanisms mediate the influence of *implicit* social cognition, giving rise to behaviour that is *more or less* subject to reflective control for which the agent is, ceteris paribus, more or less responsible. Madva summarises this point '… individuals are somewhat aware and somewhat in control of their implicit biases, and so they are somewhat responsible, and somewhat to blame' (2020: 390). The intuitively correct and generally pervasive notion of degrees of responsibility is accommodated within the model of implicit bias via increasing (or deceasing) levels of reflective input into issuing behaviour.

The main conclusions of the investigation into implicit bias and the requirements of guidance control (items 1 through to 4) are as follows:

1. An agent has guidance control as far as their deliberation mechanism is appropriately responsive to reasons (Fischer and Ravizza 1998: 41-46).

2. The appropriate way is moderate reasons-responsiveness, described in Chapter 3.
An agent's responsibility relevant mechanism K is moderately reasons-responsive iff;

(2.a) K is regularly *receptive* to reasons, (some of which are moral reasons - see item 2.a.ii).
(2.a.i) When holding fixed the operation of a K-type mechanism, the agent would recognize reasons in such a way as to give rise to an understandable pattern from the viewpoint of a third party who understands the agent's values and beliefs.
It was concluded that for implicit bias related behaviour there is an *understandable pattern*, minimally grounded in reality, in the sense that a third party could see a pattern of *biased behaviour* even though such behaviour would (perhaps significantly) be contrary to the agent's declared values. There is consistency of outcome within the integrative model of implicit bias, there is regularity, as essentially required for effective guidance control.
(2.a.ii) It is required that some of the reasons are moral reasons. It was concluded that the requirement of guidance control, that an agent recognizes reasons some of which are moral reasons, is satisfied in the case of implicit bias related behaviour.

(2.b) An agent's responsibility relevant mechanism K must also be at least weakly *reactive* to reasons; this requires that the agent would react to at least one sufficient reason to do otherwise in some possible scenario. When considered in the context of implicit bias, it was shown that issuing behaviour of implicit bias *is* moderately reason-responsiveness, in terms of being regularly receptive and weakly *reactive*, supporting the view that an agent is morally responsible for their actions.

(2.c) Guidance control also requires an agent's responsibility relevant mechanism K to be their own, where being the agent's own means 'taking responsibility' for K.
(2.c.i) This requires that the agent sees herself as the source of her behaviour. It was concluded that while seeing oneself as the source of implicit bias related behaviour is contra to the popular conception of implicit bias as essentially unconscious, the chosen model, with its element of reflective decision making and control, allows the agent to take responsibility by 'allowing an agent to see herself as the source of her behaviour'.
(2.c.ii and 3) An agent believes that she is an apt candidate for the reactive attitudes that result from how she exercises her agency in certain contexts. It was concluded that it seems reasonable to construe the reactive attitudes as potential props, part of an individual's wider strategy when taking ecological control. The condition that an agent takes responsibility by believing it is appropriate to be an apt candidate for the reactive attitudes, causes no obvious or significant problems for semicompatibilism, or guidance control, considered in light of implicit bias as characterised.
(2.c.iii) An agent must also view herself as an agent with respect to being a source of behaviour and an apt candidate for the reactive attitudes based on *evidence* for these beliefs. It was noted that taking ecological control and reflecting on evidence of personal achievement in mitigating implicit bias related behaviour encourages a sense of personal agency, of being in an important sense the source of behaviour, (and as a source, an apt candidate for the reactive attitudes).

4. Guidance control has an epistemic condition; to be praiseworthy or blameworthy an individual must know, or be reasonably expected to know, what they are doing (Fischer and Ravizza 1998: 12). John Martin Fischer's epistemic conditions that an agent must be aware of 'particular facts surrounding an action' and act with 'the proper sort of beliefs and intentions' (Fischer and Ravizza 1998: 13) were expanded to enable a more comprehensive analysis. After detailed consideration of both requirements it was concluded that the epistemic condition for guidance control was satisfied in the case of implicit bias related behaviour. For the second part of the requirement, that an agent act with 'the proper sort of beliefs and intentions' it was noted that agents recognise that actions bring about changes in the world that may legitimately produce praise or blame from others and this is for the vast majority an accepted and assumed situation. This is

true when actions are influenced by implicit bias. Implicit bias related behaviour is subject to a degree of reflection and part of an agent's character; it is reasonable that an agent recognises such behaviour, like other behaviour, brings about changes in the world that may legitimately produce praise or blame from others. The epistemic condition of guidance control is fundamental, evidenced by recurring discussion from the outset; recurring discussion also included awareness and the conscious or nonconscious nature of implicit bias.

From the summary of the investigation into implicit bias and all the requirements of guidance control (items 1 through to 4), implicit bias related behaviour is subject to guidance control and so agent responsibility. This is in harmony with the characterisation of implicit bias developed in Part II.

6.2 Semicompatibilism, Implicit Bias and Luck

In this Section I examine semicompatibilism, luck and implicit bias. Before launching into detailed discussion, it is obviously important to describe the reasons for considering luck within the context of this Thesis and outline the way forward. So far, investigation has focused on issuing behaviour of implicit bias and semicompatibilism; it was concluded that an agent *has* guidance control over issuing behaviour of implicit bias and therefore is responsible for such behaviour. I will now examine a well-known criticism of compatibilism and semicompatibilism from the perspective of implicit bias. The luck problem for compatibilism[203] has been chosen not only because it is one of the most, (perhaps *the* most), substantial problems faced by compatibilists, but importantly, implicit bias appears to be a paradigm example of a source of behaviour that is formed and often continually reinforced by factors that are to a greater or lesser extent subject to luck. Simply expressed, I have looked carefully at implicit bias and guidance control, I will now examine semicompatibilism and the luck problem in the context of implicit bias.

Drawing on *The Luck Problem for Compatibilists* (Levy 2011b), *Implicit Bias and Moral Responsibility: Probing the Data* (Levy 2017), *Semicompatibilist Options: Essays in Defense of an*

[203] One of the most incisive descriptions of the 'luck problem' for compatibilists is given by Gregg Caruso in conversation with Daniel Dennett *Just Deserts* (Dennett and Caruso 2021: 15-18). Using the luck problem, Caruso 'attacks' Dennett's compatibilist position with exceptionally clear and logical philosophical argument.

Actual-Sequence Approach to Freedom and Responsibility (T. Cyr 2018), *Moral Responsibility, Luck, and Compatibilism* (T. W. Cyr 2019), *Deep Control* (Fischer 2015c) and *Moral Luck* (Nelkin 2019)[204] I will describe how luck is claimed to threaten compatibilism and semicompatibilism. Then continue with consideration of implicit bias related behaviour *as a possible problem* for a defence of semicompatibilism from the luck problem. To be thorough, I will in turn adopt a position on implicit bias of agent responsibility for issuing behaviour and, following Levy (2017), a position that denies agent responsibility. The overall aim is to critically examine a defence of semicompatibilism from the luck problem within the context of implicit bias and so offer further comments on semicompatibilism itself.

What *is* the luck problem and what specifically is the *moral* luck problem? The Stanford Encyclopedia of Philosophy offers a concise answer: 'Moral luck occurs when an agent can be correctly[205] treated as an object of moral judgement even though a significant aspect of what she is assessed for depends on factors beyond her control' (added emphasis Nelkin 2019). This runs counter to the plausible Control Principle whereby agents are only responsible for actions over which they *have* control. Given the Control Principle and the significant contribution by luck in shaping behaviour, it is perhaps easy to become sceptical of legitimate moral assessment of almost anyone for anything. There are, of course, responses to such scepticism, some of which will be briefly mentioned, but before looking at these, it is useful for future discussion to

[204] John Martin Fischer is Chair and Dana Nelkin a member of the Dissertation Committee for Taylor W. Cyr's PhD Thesis *Semicompatibilist Options: Essays in Defense of an Actual-Sequence Approach to Freedom and Responsibility.*

[205] To say '*correctly* treated' is initially puzzling. I assume the qualifier 'correctly' simply confirms that for an agent *to be* morally lucky or unlucky she must paradoxically be, in an important sense, a 'legitimate' object of differential moral assessment, when conventional considerations of control and responsibility suggest differential moral assessment is an incorrect response. For example, when a driver unavoidably injures a pedestrian who without warning steps on to the road, it just seems a matter of fact that we do make a different moral assessment of such a driver compared to the driver who, it just so happens, makes an uneventful journey. Although I use the term 'correctly' several times in this section, it is understood that further clarification may be required.

(following Nagel 1979) quickly identify and describe four kinds of luck.[206] Resultant luck is luck in the way things actually turn out: if the *outcome* of the same intention held by two agents differs and moral assessment is correctly different, then moral luck is in play within the process of assessment. An example will illustrate the general idea. Two conscientious people (genuinely) forget to check their brakes and then experience brake failure. In one case a pedestrian without warning steps on to the road and is injured because the approaching car cannot stop in time, in the other case it just so happens that no pedestrian steps out and the second driver completes their journey without incident. Differential moral assessment is a case of moral luck. One driver will be the target of considerable blame and resentment, the other driver simply completes their journey without incident, yet all that separates them is an unlucky event, not within the control of either driver. Circumstantial luck or present luck is, as the name suggests, luck present within the circumstances in which agents find themselves that is beyond their control; through bad luck, being in the wrong place at the wrong time, (or good luck contributing to being in the right place). To make correct differing moral assessments is a case of circumstantial moral luck. Constitutive luck concerns luck in who we are, our attitudes (questionable), disposition, and so on. This includes who we are because of our genetic inheritance and (to a greater or lesser extent) our environment. Agents clearly do not control their genetic inheritance yet correctly blaming for behaviour that in some way originates from an agent's genetic makeup is a case of constitutive moral luck. Finally, causal luck concerns something familiar, the notion that current actions are the outcome of causal links over which an agent has no control and hence, following the Control Principle, is not responsible. Differing moral appraisal of behaviours that originate in these ways are examples of moral luck.

The above touches on several types of luck that arguably diminish or remove agent control and so responsibility. Yet, differential moral assessment, as the example above suggests, is quite commonplace in everyday life. Nelkin (2019) notes, there are three general approaches to the problem of moral luck. First, to deny that moral luck

[206] Paul Russell *Freewill Pessimism* (Russell 2015a) describes very clearly the four forms of luck identified by Nagel and discusses their control limiting consequences that suggest 'since nothing is properly and fully under the agent's control, there are no suitable foundations for moral evaluation or moral responsibility' (Russell 2015: 5).

exists, even though appearances suggest otherwise. Second, accepting that moral luck exists while rejecting or restricting the control principle. Third, the idea that it is incoherent to accept *or* deny the existence of some forms of moral luck, particularly with constitutive luck.[207] It is beyond the scope of this Thesis to discuss these responses to moral luck in detail, rather, I will now move directly to careful consideration of the threat posed to semicompatibilism by luck, specifically circumstantial and constitutive luck.

The essential problem for compatibilists such as John Martin Fischer, who's semicompatibilism is largely built on the claim that 'moral responsibility is an essentially historical notion' (1998: 170), where agents over time take responsibility for their actions by *owning* the mechanism that is the source of their actions, is the very nature (constitution) *of* the mechanism may be significantly subject to constitutive luck, throwing into doubt the sense that the mechanism is truly *mine*. (Also recall the similar regress problem associated with Kane's 'self-forming actions' and related notion of 'will-setting actions' page 41). A semicompatibilist may argue that constitutive luck can only excuse for so long, because over time an agent has the possibility to take responsibility for dispositions and values acquired through constitutive luck. On this view, agents reject some dispositions and accept others; an agent creates something of their *own* from what they are given. John Martin Fischer makes a similar point, using the expression 'playing the cards that are dealt you' (2015c: 164). (I will outline Fischer's response to the luck problem in the concluding remarks of this section). However, *if* an agent's mechanism is in *any* sense a function of luck, then to that extent it is not under their control and so issuing behaviour is not entirely controlled and responsibility is diminished. Levy expresses the idea succinctly, 'the problem of history *is* a problem of luck' (2011b: 87), alternatively, Dennett argues 'Strawson may have said that "luck swallows everything" but if so he was wrong. Luck sets the stage …' (Dennett and Caruso 2021: 20). Recalling earlier discussion of the role of culture in forming attitudes, beliefs and biases, it is easy to see how implicit bias engages with the luck problem in the sense of circumstantial and constitutional luck; the formation of biases, their nature and degree, depends (to greater

[207] See also Levy, *Hard Luck: How Luck Undermines Free Will and Moral Responsibility,* Chapter Two, An Account of Luck (Levy 2011a) for discussion of this point, the four forms of luck mentioned above and a clear introduction to luck.

or lesser extent) on circumstances and an agent's constitution, both (again, to a greater or lesser extent) matters of luck and, it is argued, beyond their control.

Criticism of the plausible idea that agents can take responsibility via an historical process of *deliberation* that mitigates, perhaps effectively eliminates, the role of luck in the formation of their endowment issuing behaviour, is clearly described by Levy as the 'luck pincer':

> In so modifying themselves, they satisfy the most demanding history-sensitive compatibilist ownership conditions. But the series of decisions whereby they supposedly took responsibility for their endowment were either expressive of that endowment, or subject to present luck, or expressive of their endowment as modified by previous actions in turn expressive of their endowment or subject to present luck. The solution to the problem of constitutive luck is therefore lots more luck. But surely, we cannot undo the freedom-undermining effects of luck by virtue of more luck. Iterated luck does not cease to be luck. (2011b: 96)

Levy's argument is important and persuasive; drawing on Cyr (2018) I will shortly look at a response to Levy. Before this, an important comment concerning implicit bias and luck. The luck pincer problem, whereby an agent's deliberation cannot be free from the influence of luck on formation of their endowment and present circumstances, suggests a problem with the adopted model of implicit bias. Consider the model of mechanisms that mediate the influence of implicit social cognition (Fig 5.1), the model adopted to characterise implicit bias. I have previously argued the presence of deliberation and reflection within the model facilitates responsibility for issuing behaviour. However, if luck in various forms is present, unavoidable and influential during deliberation and reasoning when challenging acquired values and biases or making decisions about behaviour, then agent responsibility *for* issuing behaviour is clearly and perhaps significantly brought into question. The luck problem is therefore threatening to semicompatibilism *and* the characterisation of implicit bias so far presented; what response can be made to this challenge?

Anticipation of Cyr's response to elements of this threat may be found in Judith Andre's short Paper *Nagel, Williams, and Moral Luck* (1983). In contrast with Kant,[208] (that

[208] A clear exposition of the contrasting positions of Kant, Adam Smith, Bernard Williams and others is given by Simon Blackburn during his lecture 'Moral Luck and the Peculiarity of Blame' at The New

everyone as rational agents can act morally no matter what their circumstances), Andre describes Thomas Nagel (1979) and Bernard Williams' (1981) position, that 'in practice we evaluate actions and agents *partly* on the basis of circumstances beyond the agent's control - on the basis of luck.' This leads to '… inconsistency and possibly incoherence within the concept of morality.' Also, '… Williams (argues) that rational justification rests *partly* on luck' (Andre 1983: 202). I have emphasised 'partly' as this, I believe, is the key notion employed by Cyr in his defence of compatibilism from the luck challenge.

Recall, that it is claimed that an agent's endowment, their set of values, dispositions and biases, to a greater or lesser extent, are subject to constitutive luck. Given that part of the process that leads to action is subject to constitutive luck and not under control by the agent, then to this extent an agent is not responsible for their actions. Due to the luck pincer, an agent's endowment is claimed to be a matter of luck for them but following a compatibilist account such as John Martin Fischer's semicompatibilism, an agent *can* assume responsibility over time by progressively owning their endowment. As noted, Levy does not accept that ownership over time is possible because of the luck pincer, the *constant* ongoing influence of present luck. Cyr disagrees, arguing that while agents may *sometimes* be presently lucky, 'it is not the case that they are presently lucky in every case of taking ownership for their endowments' (2018: 60). (Dennett makes the same point, claiming the actions we perform are not entirely dependent on luck. (Dennett and Caruso 2021: 20)).

Cyr's comprehensive defence of compatibilism from present and constitutive luck employs one essential tactic: the notion of degrees of responsibility. Cyr argues that agents are not presently lucky on *every* occasion that an agent builds their endowment, so taking ownership *is* ultimately possible and involves agent reflection on their values and dispositions. While agents are not completely immune from the influence of present luck, there is partial responsibility for how the agent's endowment, their values and beliefs, evolve over time. On this view, how an agent is, their constitution, is not wholly a matter of luck and issuing behaviour is to that degree the responsibility of the agent.

College of the Humanities on the 26th February 2016, available on YouTube (https://www.youtube.com/watch?v=bRaRVx_7tVY).

Cyr gives further explanation of his defence, beginning with his expression of the 'pincer problem':

> How is it that an agent can be responsible for the evaluation and adjusting of her endowment, when the toolkit, so to speak, is a part of her very endowment? If the agent is responsible for taking ownership of her endowment because her history includes previous modifications to her endowment, then the problem is not solved but merely moved back in time to prior uses of the toolkit with which she has been endowed. (Cyr 2019: 13)

How can an agent 'break into' this closed system to take ownership? Cyr does not believe the system *is* completely closed, arguing by example that on the occasion of an agent's first potentially free and morally responsible action, although the agent may up to that point be entirely constitutively lucky, she can *nevertheless* be *slightly* responsible for her action, be a 'little agent' (following Cyr 2019: 13). The plausible notion of degrees of responsibility is employed to begin an explanation of how agents can, over time, begin to take a degree of ownership of their endowment even considering constitutive luck. It seems that at the point of being slightly responsible for her action, within Cyr's idea of a 'little agent' is the notion of a 'flicker' of (some form of) agent causation that allows overtime increasing ownership by the agent of their endowment, not entirely defined by luck. (Appendix B describes in greater detail the concept of agent causation).

Having looked briefly at luck, how it challenges responsibility and a brief outline of a defence of agent responsibility considering present and constitutive luck, I will turn to implicit bias; the soundness of Cyr's defence of compatibilism against luck will be assessed considering two essentially different perspectives on implicit bias.

As noted, the formation and influence of implicit bias appears a paradigm example of how luck shapes behaviour. Is there something *special* about implicit bias, as part of an agent's overall endowment of beliefs, opinions and attitudes that calls for special consideration when thinking about defence of compatibilism from the luck problem? Is the overarching problem of luck for moral responsibility particularly acute or in some way unique for implicit bias influenced behaviour? Although it seems reasonable and plausible to say that Cyr's approach to the problem of luck is relevant when implicit bias is included within the overall endowment of an agent, clearly this must be looked at carefully. First, when issuing behaviour of implicit bias is subject to a degree of rational deliberation and subject to guidance control and responsibility. Second, when agent responsibility for behaviour is absent. The first conception of implicit bias has been described at length, but the second conception needs more explanation, and I will revisit Levy's Paper *Implicit Bias and Moral Responsibility: Probing the Data* (2017) to give further insight, before investigating if these perspectives on the nature of implicit attitudes impact Cyr's defence of compatibilism from the problem of luck.

Levy (see also, page 137) says in cases in which an action (or its consequences) has a moral character due to the agent's implicit attitudes, control over the action (or its consequences) is greatly diminished to such an extent that it is highly plausible the agent lacks responsibility-level control:

1. Moral responsibility requires that an agent exercises responsibility-level control over their action or the consequences of their action (depending on whether they are putatively responsible for the action or for its consequences).
2. In cases in which an action (or its consequences) has a moral character due to the agent's implicit attitudes, control over the action (or its consequences) is greatly diminished.
3. The decrease in control is significant enough to make it *highly plausible that the agent lacks responsibility-level control.*
4. If control is a necessary condition of direct moral responsibility, agents are therefore not responsible for these actions or their consequences (added emphasis 2017: 6).

This is based on the view (generally contra to that expressed within this Thesis) that the processes that mediate implicit attitude influenced behaviour are not available to introspection; 'we lack insight into what influence they (implicit attitudes) have on our perceptions and judgments, and there are no reliable means of modulating or inhibiting this influence' (2017: 8). That said, Levy continues by looking more critically at the type of control of actions that is necessary to realise responsibility-level control; even if control over actions is diminished, is such control still sufficient for agent responsibility? Following John Martin Fischer, Levy begins by discussing 'understandable pattern(s) of reasons recognition, minimally grounded in reality' (Fischer and Ravizza 1998: 73). Levy suggests that sensitivity to reasons is patterned when it is continuous, broad and systematic[209] and summarises his control condition sufficient for agent responsibility as follows: if an action is caused by mechanism m, and m is sensitive to reasons in a suitably patterned way, then m realizes control over that action (2017: 11). Having set up this control condition, Levy then looks at the impact on reasons-responsiveness when an implicit attitude is a part of the mechanism. Summarising Levy's developing ideas: an associative account entails that implicit attitudes will display little reasons responsiveness, hence there is insufficient control over the moral character of our actions, (when this character is due to implicit attitudes), for us to be morally responsible for their having that character (following closely Levy 2017: 12). By contrast, as previously mentioned, (page 108), Mandelbaum considers implicit attitudes *are* unconscious beliefs. On this view, mechanisms that have implicit attitudes as components will not show *any* reduction in reasons-responsiveness. Levy does *not* a accept implicit attitudes are beliefs, claiming that while they have some propositional structure, they do not have the right kind of structure to underwrite continuous, broad and systematic responsiveness. Rather, they have a 'patchy propositional structure, not the kind of continuous and broad propositional structure we rightly associate with beliefs' (Levy 2017: 13). Levy presents various examples to show that such patchiness undermines patterned reasons-responsiveness, control and responsibility. However, recognising that undermining responsibility does not sit well in many cases where there is insensitivity to reasons, Levy considers, and under specified conditions accepts, an alternative to the control condition

[209] 'To say that it is *continuous* is to say that the relevant mechanism is sensitive to relatively fine-grained alterations in the parameters of a particular reason. The violinist exhibits continuous sensitivity to orchestral dynamics, say, when she would respond not merely to *some* alteration in the volume of the orchestra by adjusting her own volume, but when she is appropriately responsive to a (relatively) continuous dynamic range. To say that sensitivity is *broad* is to say that the relevant mechanism is responsive not just to a particular kind of reasons (however continuously) but to a range of different kinds of reasons. The violinist shows exquisite control since she is sensitive not only to orchestral dynamics, but to the acoustics of the room, the mood of the conductor and the audience, and so on. Sensitivity is *systematic* when the mechanism would respond to a particular kind of reason in any context (so long as that context does not include features that neutralize the reason). The violinist exhibits systematic responsiveness to dynamics when she would respond to them in a large hall or a small, with a full orchestra or a small ensemble, and so on' (Levy 2017: 10).

approach, referred to as the attributability view. On this view, 'agents are responsible for attitudes that properly belong to them, and for actions caused by such attitudes' (Levy 2017: 18).

The attributability position may be summarised as follows: Control is not a *necessary* condition of moral responsibility but may be relevant, and when it is relevant it is 'because actions that are controlled by agents are typically attributable to the agent in a way sufficient to ground their responsibility for the action' (Levy 2017: 18). If agents fail to exercise responsibility level control, (for example, because of implicit attitudes), they may still be responsible *if* implicit attitudes are deeply enough attributable to the agent to ground responsibility. This is essentially the attributability position; for implicit attitudes, agents are *not* responsible for issuing actions or their consequences:

1. Agents are morally responsible for actions or the consequences of their actions when they are caused … by attitudes that are sufficiently deeply attributable to them.
2. In those cases in which an action (or its consequences) has a moral character due to the agent's implicit attitudes and would lack that character were the action controlled by their explicit attitudes, the attitude is not deeply attributable to the agent.
3. If attributability is a necessary condition of *direct* moral responsibility, agents are therefore not responsible for these actions or their consequences (Levy 2017: 19).

The *key* claim here is that implicit attitudes do not belong to an agent's deliberative standpoint, an explicit and considered mesh of consistent beliefs and attitudes in a way that is sufficient for responsibility level attributability. However, Levy reminds us that implicit attitudes are patchy endorsements, and as such are not *entirely* disconnected from an agent's deliberative standpoint. Having links with other attitudes that are firmly attributable to the agent suggests such links can afford *some* degree of attributability of implicit attitudes to an agent.[210]

Levy does not share the conclusion adopted within this Thesis, that agents are responsible for implicit bias influenced behaviour, however, Levy's two perspectives on implicit attitudes and responsibility, the control condition and the attributability view, are generally in accord with the model of implicit bias I have adopted; the interactive

[210] Recall Holroyd and Kelly's view (page 138) that even *if* it is accepted that implicit associations are *not* unified within an agent, this does not entail implicit associations contribute nothing to who the agent is or cannot be to that extent subject to evaluation.

model, being a configuration that includes decision making and implicit social cognition components that are responsive to reasons and Holroyd and Kelly's (2016) approach, where implicit bias is considered part of a person's character.

In contrast to the model of implicit bias I have adopted within this Thesis, while noting the qualification above, (that links with other attitudes that are firmly attributable to the agent suggests *some* degree of attributability of implicit attitudes to an agent), Levy answers the following questions as follows (Levy 2017: 6):

1. Do agents exercise a sufficient degree of control over actions that have a moral character due to their implicit attitudes to appropriately be held morally responsible for them? Levy answers no.
2. Are actions with a moral character due to implicit attitudes caused by states that properly belong to the agent, (and so appropriately be held morally responsible for them)? Levy answers no.

To summarise, two positions on implicit bias and agent responsibility have been described. First, a position developed earlier and adopted within this Thesis, where the nature of implicit bias motivates agent responsibility for issuing behaviour. It has been argued that such behaviour *is* legitimately subject to guidance control, hence freely and responsibly undertaken as described by semicompatibilism. Second, a position where the nature of implicit bias *does not* motivate agent responsibility for issuing behaviour, summarised by Levy:

> Implicit attitudes do not seem properly to belong to agents' deliberative standpoints, in the way required for responsibility-level attributability. The conditions under which implicit attitudes come to be suitably annexed to the agent's deliberative standpoint are conditions under which it is no longer true that their actions have a moral character due to their implicit attitudes. So, if control, or attributability, or both, are necessary conditions of moral responsibility, agents are not directly responsible for actions that have a moral character due to their implicit attitudes. (2017: 26)

It is beyond the scope of this Thesis to evaluate these differing positions; the overall undertaking is ultimately to look critically at semicompatibilism, not implicit bias. As noted, an alternative perspective on implicit bias has been introduced with the intention of greater thoroughness in what is to follow. Semicompatibilism is at risk from the problem of luck; the next task is to look at how Cry's plausible defence of semicompatibilism from the luck problem stands up considering both implicit bias perspectives. If Cyr's defence falters under these circumstances, then a weakness in the semicompatibilist position with respect to luck is exposed, (or some aspect of Cyr's defence requires attention).

The shape of this enquiry is as follows: Summarise briefly the essential problems that luck causes with respect to (semi)compatibilism and the key elements of Cyr's defence, then carefully check the elements considering implicit biases forming part of an agent's endowment or mechanism. How is the defence argument affected when implicit bias is introduced? Each implicit bias model will be considered, beginning with the most familiar, developed earlier within this Thesis, where the nature of implicit bias motivates agent responsibility for issuing behaviour.

Summarising the discussion of luck and compatibilism: The most important feature of luck is its threat to moral responsibility due to related lack of control. In addition, to be lucky for an agent, an event or situation must have some importance to them; I have no control over London traffic, but it is of no importance to me that it is running particularly smoothly because I am currently working in Wales, therefore, such traffic conditions are not a lucky situation for me. Further, a lucky action or event is one that does not occur in a large proportion of nearby worlds (Cyr 2018). Luck threatens agent control over who they are, (and their issuing mechanism), and so their responsibility for who they are and what they do. History sensitive compatibilists, while accepting constitutive luck in the formation of agent endowment claim that agents *can* take control and be responsible for their endowment by taking ownership of their values and attitudes over time. Present day luck threatens to upset the possibility of taking ownership and so control and responsibility: Agent deliberation that should lead to ownership is subject to *present day luck*, hence the aim of taking control over time is defeated. As described, Cyr's essential counterclaim is that agents are not presently lucky *every time* they take ownership of their endowments; moral education, self-discipline, and so on, *can* lessen the influence of present-day luck sufficiently for an agent to take ownership of their values and mitigate constitutive luck. However, moral education and self-discipline are subject to constitutive luck, suggesting a regression problem likely to undermine taking ownership, but examples (Cyr 2018: 73) claim to show a *small degree of moral responsibility* is plausibly present from an early age; from a starting position of 'a small degree of moral responsibility' an agent's moral responsibility is claimed to increase overtime, i.e., is a function of their history. Essentially, present day luck is not all pervasive, the agent has various means at their disposal to mitigate its affects and from a starting position completely dominated by constitutional luck an agent *can* take moral responsibility overtime.

Having summarised the essential problems that luck causes with respect to (semi)compatibilism and the key elements of Cyr's defence, does the presence of implicit bias, within an agent's assortment of implicit and explicit attitudes and beliefs, cause any problems for Cyr's argument? The implicit bias model adopted within this Thesis, where behavioural decisions are subject to a degree of reflection, does not in itself reveal any issues for Cyr's defence of compatibilism from the luck problem. Recall, from Chapter 5 it was concluded that while greater understanding is necessary concerning the *reflective* and impulsive processes that mediate between implicit social cognition and explicit behaviour within the interactive model (Deutsch and Strack 2010: 73), there is no doubt that control in an important sense, and so responsibility, is possible within this model. Cyr defends the *possibility* of meaningful reflection from the luck problem; the presence of implicit bias within an agent's mix of attitudes and beliefs does not of itself affect the defence argument because by their nature implicit biases are included within the scope of agent reflection and deliberation.

Consider now, the characterisation of implicit bias described earlier during discussion of Levy and the attributability position. Does the presence of this interpretation of implicit bias within an agent's endowment cause any problems for Cyr's argument? This is a more demanding situation to analyse. Some key elements of this characterisation of implicit bias are given below, drawn from Levy's Paper *Implicit Bias and Moral Responsibility: Probing the Data* (2017).[211] The list is not comprehensive, but based on relevance to the discussion to follow:

[211] I have reformatted and made minor changes to Levy's text to aid presentation while hoping to retain and not diminish Levy's intended meaning, see (Levy 2017). See also *Consciousness and Moral Responsibility* (Levy 2014b) and *Précis of Consciousness and Moral Responsibility* (Caruso 2015); Levy's Paper is key within the debate, developing a 'consciousness thesis' and 'global workspace theory' that are employed to

1. Mechanisms with implicit attitudes as components may show complete insensitivity to particular reasons and thereby cause actions that have a moral character they otherwise would not have had.

2. The *deliberative standpoint* is a test for the degree to which the attitude belongs to the agent. Attitudes are acquired through engagement with the deliberative standpoint only when they have appropriate relations with those that make up the standpoint; conversely, attitudes acquired in ways that *bypass* the deliberative standpoint will not become enmeshed in it. Of course, we can apply this test only if we can confidently identify some attitudes as properly belonging to the deliberative standpoint. We can often identify such attitudes by way of their role in the behaviour of agents: insofar as they are implicated in consistent and instrumentally rational decision-making and behaviour, these attitudes may be said to *partially* constitute the agent's deliberative standpoint.

3. There is plentiful evidence that implicit attitudes can be acquired in ways that *bypass* and even conflict with the attitudes that constitute an agent's standpoint.

4. Evidence about how agents acquire implicit attitudes is evidence for the degree to which such attitudes belong to their deliberative standpoint because under ideal conditions, agents acquire and maintain attitudes only when they are consistent with the attitudes constitutive of that standpoint; inconsistency should lead either to revision of their former attitudes or rejection of the new attitude. Evidence that agents acquire attitudes that are inconsistent with their (continuing) attitudes is therefore evidence that the attitudes acquired *do not* belong to the agent. Similarly, recognition of an attitude's inconsistency with other attitudes should lead to the elimination of one or other: if both persist, we have reason to think that one or the other should not be fully attributed to the agent. (When acquired implicit attitudes *are* consistent with continuing attitudes, this is surely indicative of implicit attitudes belonging to the agent with related responsibility; for example, a situation where a person is both explicitly and implicitly biased).

critique two leading accounts of necessary conditions for moral responsibility, real self-accounts and control-based accounts. Levy concludes that implicit bias is 'not plausibly taken to be an expression of [the agent's] evaluative agency, their deliberation and evaluative perspective on the world' therefore based on the real self account an agent is excused moral responsibility. Similarly, on control-based accounts, *including* Guidance Control, Levy argues that agents are not responsible for implicit bias influenced behaviour (2014b: 95), (2014b: 115).

The intuition is that implicit attitudes when characterised in this way are, in an important sense, 'outside' of an agent's deliberative standpoint (item 2, 3 and 4), therefore on this view not available as an object of rational reflection (item 1). As noted previously, Cyr's defence of compatibilism from the luck problem requires the possibility of meaningful reflection; the *absence* of implicit bias from an agent's mix of attitudes and beliefs that make up their deliberative standpoint reduces significantly, perhaps eliminates entirely, the possibility of rational scrutiny. Implicit biases, on this account, are subject to ongoing constitutional luck with attendant threat to agent responsibility.

So, Cyr's defence works if implicit attitudes are included within an agent's deliberative standpoint. Levy's alternative characterisation of implicit bias takes such attitudes outside of an agent's deliberative standpoint, unavailable to Cyr's defence, hence subject to constitutional luck. Compatibilism is still threatened by the luck problem in the context of agent implicit bias characterised by Levy. It has been shown that Cyr's defence of compatibilism falls short, in that it leaves (semi)compatibilism exposed to the luck problem when considered in the context of Levy's credible account of implicit attitudes. The term 'falls short' is used because *other* attitudes, beliefs and dispositions *are* included within an agent's deliberative standpoint and so defended from the luck problem by Cyr's argument. Further, it can be noted that based on this understanding of implicit attitudes, Levy concludes 'if control, or attributability, or both, are necessary conditions of moral responsibility, agents are not directly responsible for actions that have a moral character due to their implicit attitudes' (2017: 26); a position contra to that adopted within this Thesis, based on a different characterisation of implicit bias.

Before drawing final conclusions concerning semicompatibilism, implicit bias and luck there is a further point to make. While it can be seen how luck may affect the composition of an agent's endowment, the very elements that an agent exercises to take ownership *of* their endowment, there is an intuition, an internal sense, that intervention into this process or some other consideration may be available. Earlier discussion of Kant and Freud (pages 23 and 27) perhaps motivate this intuition when it was noted that for Freud, reason has a special standing, having the capacity, (through psychoanalysis), to be free from the influence of desires, (unconscious forces), and act *autonomously* in accordance with reason alone and moral duty. Or with Kant, where reason is independent of the natural world of appearances, causation and luck, directing human beings freely and autonomously within the moral landscape. Margaret Walker (1993) asks us to imagine what it would be like to live among what she describes as 'pure agents', where following Kant, this refers to agents whose reason is independent of the natural world of appearances, causation and luck. Walker's argument is detailed and subtle, but the essential point is that pure agents are not burdened with blame by others for matters that are the outcome of luck, nor do they engage in blaming others where an unlucky situation has occurred. It is clearly difficult to imagine such a world, but an absence of blaming or being blamed for an unlucky state of affairs at first look appears quite attractive. However, if praise and blame are absent from situations thought to be the result of luck then perhaps other responses could also be thought inappropriate, such as sympathy, charity and empathy. Such an absence of personal response to difficulties that are the consequence of luck is surely undesirable. Walker argues that the 'impure agent', an understanding of agency that *incorporates* moral luck, 'is not the worst we can do' (1993: 247). A purely rationalist approach, where blameworthiness is purely a function of faultiness, for example, if both drivers carelessly do not check their brakes and, on this account, are *equally* at fault when one driver thereby causes a fatality due to an unlucky event, is certainly counter intuitive. There is clearly something wrong or missing. Susan Wolf *The Moral of Moral Luck* (2001) develops a position that incorporates the rationalist position of blameworthiness as a function of faultiness, *plus*, following what is referred to as an irrationalist position, acknowledgement and approval of different emotional responses to different *outcomes*, from others and the agent who brought about the effect or outcome (Wolf 2001: 13). For the agent, feelings of guilt or regret are wholly appropriate and rightly proportional to the severity of the outcome. Referring to the example above, feelings of regret would clearly be present and appropriate if the brake failure element is removed and there was no faultiness at all; a pedestrian steps out such that the driver of a car with excellent brakes just cannot stop in time. These brief outlines of further reactions to the problem of luck, in terms of the 'impure agent' and the 'irrationalist position', are given to illustrate the diversity of response to the problem of luck.

John Martin Fischer addresses the big issues of luck and the source of moral responsibility in Chapter Ten of *Deep Control* (2015c). After careful and detailed consideration, the key conclusion is:

> The mistake is to suppose that compatibilism seeks to identify an 'Island of Control' — an Inner Citadel. It is better to think of compatibilism as conceding from the beginning that we are *thoroughly subject to factors entirely outside our control.* Nevertheless, according to the compatibilist, we can still exhibit a meaningful and robust sort of control. It is not as if the compatibilist seeks to carve out a sphere of pure 'internality' and immunity to arbitrariness, and then must be embarrassed to discover that the inner sanctum is not secure. He never thought that we needed such a place. (added emphasis 2015c: 182)

The chapter from which the above quotation is taken concerns 'sourcehood', examining the notion of 'ultimate control' or 'self-creation'. Using several examples, John Martin Fischer argues such ultimate control is *too* demanding and accepts that we are thoroughly subject to factors entirely outside of our control and yet still retain agency and responsibility enabling control, '*in a suitable sense*' (added emphasis 2015c: 185). We can be, using Fischer's example, accountable for playing the cards that we are dealt, even if we did not make the cards or invent the rules of the game; we can be responsible without *ultimate* control or self-creation. In this Paper, John Martin Fischer meets head-on the challenges faced by semicompatibilism and guidance control due to luck, agency, determinism and the notion of ultimate control; the essential conclusion is that meaningful and responsible behaviour *is* possible and demands for ultimacy are essentially meaningless.

In summary, luck is often claimed to be the most serious threat to historical notions of (semi)compatibilism by taking away vital responsibility enabling control. A defence of semicompatibilism from the luck problem was described based on the essential claim by Cyr (2019) that an agent is not presently lucky all the time, and an agent has the possibility of mitigating the effects of luck by employing self-control and drawing on their moral education. A regression problem is avoided by the notion of a 'little agent'; an agent's first potentially free and morally responsible action, although *at that point* entirely constitutively lucky, can *nevertheless* be *slightly* responsible for their action (following Cyr 2019: 13). The defence of semicompatibilism was reviewed considering two models of implicit bias, paradigm sources of behaviour that are subject to constitutive and present luck. First, the model adopted within this Thesis, where, simply expressed, agents are responsible for implicit bias issuing behaviour and second, following Levy (2017) a model where agents are not responsible. It was concluded that Cyr's defence is not affected considering the first model when implicit attitudes are included within an agent's deliberative standpoint. Levy's alternative characterisation of implicit bias takes such attitudes outside of an agent's deliberative standpoint, unavailable to Cyr's defence, hence subject to constitutional luck. Compatibilism remains threatened by the luck problem in the context of agent implicit bias characterised by Levy and with respect to Cyr's defence. This section concluded with some thoughts with reference to Kant (page 23) and Freud (page 27) concerning independence of reason from the material world, including luck, and a statement of John Martin Fisher's position on luck and sourcehood.

6.3 Summary of Part III

Chapter 6 examined semicompatibilism/guidance control and a defence of semicompatibilism from the luck problem in light of implicit bias. The essential conclusions are these; implicit bias related behaviour was shown to be subject to guidance control and so agent responsibility. This is in harmony with the models of implicit bias developed in Part II, in the sense that semicompatibilism/guidance control *and* the models of implicit bias support agent responsibility for behaviour influenced by implicit bias. No problems with semicompatibilism/guidance control were found considering implicit bias related behaviour. An assumption, I believe to be reasonable, was made during this investigation concerning John Martin Fischer's position on the epistemic condition of guidance control.

Having shown that implicit bias does not generate problems for semicompatibilism and guidance control in conferring responsibility level control of behaviour, I looked at implicit bias and related behaviour as a potential difficulty for a particular defence of semicompatibilism from the important luck problem. It was concluded that Cyr's defence is not affected when considering the first model of implicit bias when implicit attitudes are included within an agent's deliberative standpoint. Levy's alternative characterisation of implicit bias takes such attitudes outside of an agent's deliberative standpoint, unavailable to Cyr's defence, hence subject to constitutional luck. Compatibilism is still threatened by the luck problem in the context of agent implicit bias characterised by Levy. This is clearly important, as a defence of semicompatibilism from perhaps its most significant threat is jeopardised if Levy's understanding of implicit bias is correct.

Finally, on page 201 it was noted that the luck problem is … threatening to semicompatibilism *and* the characterisation of implicit bias so far presented. *Why* is luck a problem for the characterisation of implicit bias with respect to the integrative model? The issue here is far from unique, in the sense that it is an example of general pervasive luck influencing construal of situations, formation of attitudes and deliberation itself. A possible response to this problem could generally follow Cyr, by developing an argument that influence of luck within the interactive model is *not* all pervasive and there is room to build sufficient autonomous reflection that warrants responsibility enabling control.

Holroyd and Kelly's perspective, described in *Implicit Bias, Character, and Control* (2016), where implicit biases are part of who the agent is and agents can be evaluated for being influenced by them because the agent has control (i.e., ecological control) over such mental entities (2016: 175), does not consider moral luck. The implications of 'the luck problem' for Holroyd and Kelly's perspective[212] on implicit bias and responsibility cannot be explored here, however, much of what has already been considered in relation to luck is relevant. Following the comments above it is worth recalling that while it is claimed that the luck pincer, the constant ongoing influence of present luck, is problematic for ownership over time of our character, there are many notable voices that disagree, arguing that while agents may *sometimes* be presently lucky, 'it is not the case that they are presently lucky *in every* case of taking ownership for their endowments' (Cyr

[212] Although clearly within a different context, (deprivation, as a threat to the state's legitimate punitive authority), it is clearly of great important to note Holroyd's description of the role of luck or more likely bias *Punishment and Justice* (2010): 'Further, we can append the example (of the impoverished parent) to make clear that the circumstances of want (as is the case in many instances of disadvantage or deprivation) result not from foolishness or poor motivation on the part of the disadvantaged, but rather from *bad circumstantial luck*, or *more likely*, a cumulative and *pervasive infrastructural (and sometimes overt and personal) bias*. We need only to consider the data that report that, for example, in the U.K. around two-fifths of people from ethnic minorities live in low-income households—twice the rate for white people—to see the plausibility of this assumption. That disadvantage tracks identity traits such as race, gender, and class should confirm the injustice it entrenches' (Holroyd 2010: 94).

2018: 60). As previously noted, Dennett makes the same point, claiming the actions we perform are not entirely dependent on luck. (Dennett and Caruso 2021: 20)).

Conclusion

> *Choice and consciousness are one and the same thing.*
> Jean-Paul Sartre[213]

An early interest in Sartre's *Roads to Freedom*[214] led to wider reflection on human freedom and responsibility, and ultimately to the question addressed by this Thesis; does implicit bias threaten the semicompatibilist position on free will and responsibility?

It was important to place this question in the larger historical context of the free will debate. Part I concluded with an account of John Martin Fischer's free will semicompatibilism to be taken forward into Part III. Part II had a similar purpose, seeking a clear position on implicit bias, particularly control of and responsibility for implicit bias influenced behaviour.

Establishing a clear position on implicit bias was challenging, not least because of the many and sometimes conflicting views[215] about what implicit biases are, how they are acquired, our possible awareness of and responsibility for related behaviour and whether mitigating strategies work. The purpose was not to present a definitive and conclusive argument for a particular position on these issues or make a comprehensive comparative study, but present a clear, defensible and plausible choice, one supported by substantial credible theory and practice. A model meeting these requirements was chosen based on Deutsch and Strack (2010) together with the contrasting approach of Holroyd and Kelly (2016). Importantly, both approaches conclude there is individual responsibility for behavioural expression of implicit bias. This is a key point; I have examined semicompatibilism considering a model of implicit bias where agent

[213] *Being and Nothingness* (Sartre 1984: 595).

[214] The first episode of *The Roads to Freedom* was broadcast on Sunday, 4th October 1970. The series was repeated on television once in 1976 and then disappeared completely for 36 years until a one-off screening at the BFI in 2012. Although there is great demand for the series to be made available on DVD, or simply broadcast again, there has been nothing since the 2012 showing.

[215] '…many and sometimes conflicting views…' clearly, also describes the free will debate!

responsibility for implicit bias related behaviour is acknowledged, from two substantial and different perspectives.

Four questions may be raised by this approach: Was it a reasonable and promising idea to employ implicit bias as a means of critically assessing semicompatibilism, and why were two implicit bias perspectives chosen? How important was the choice of implicit bias model? If the choice of implicit bias model had a significant influence on how semicompatibilism responds, then how meaningful was the critique?

I have argued that implicit bias is a valuable[216] and previously unexplored perspective from which to conduct a critique of semicompatibilism. (As previously noted, Brownstein makes an encouraging comment that implicit bias is '… a good test case for theories of moral responsibility that aim to accommodate the messy reality revealed by contemporary sciences of the mind' (2016a: 766)). Initially, a wholly unconscious understanding of implicit bias and influenced behaviour strongly suggested an absence of agent control and so responsibility, even though it felt inappropriate that such behaviour should escape moral appraisal. Careful investigation of implicit bias showed strong and plausible arguments and evidence *supporting* agent responsibility for issuing behaviour, but would John Martin Fischer's semicompatibilist model respond to the complex phenomenon of implicit bias with endorsement of agent responsibility? This is now familiar territory, but I believe it is worth confirming the importance and legitimacy of bringing together in a critical way the current and vital issue of implicit bias and a major position within the longstanding free will and responsibility debate. Two implicit bias perspectives were chosen; one essentially a psychological approach, an interactive model based on the work of Deutsch and Strack (2010), the other a philosophical approach by Holroyd and Kelly (2016).[217] Two approaches were chosen with the intention of presenting a robust and defendable position to take forward into Part III; two different ways of understanding implicit bias, both granting agent responsibility for behaviour. This kind of starting position is reflected in the first section

[216] As noted in the Introduction to this Thesis, implicit attitudes can be directed toward many things, but it is the 'very morally weighty' judgment and evaluation of existing stereotypes or stigmatized groups *of people*, and such discriminatory behaviour generally, which makes implicit bias *matter* (following Brownstein 2016a: 765).

[217] See also Jules Holroyd *What do we Want from a Model of Implicit Cognition?* Proceedings of the Aristotelian Society, Issue No. 2, Volume CXVI (2016).

of Vargas *Implicit Bias, Responsibility, and Moral Ecology* (2017); ' … the action must be suitably related to some internal feature of the agent, that is, some bit of psychology arranged in this way rather than that, such that the agent identified with the action, or that it flows from the agent's values, or that it was a product of the agent's rational or normative capacities' (2017: 220). These approaches suggest various positions on agency, and it would be valuable to explore in some detail these models, and others, from this perspective. (See Appendix B Agent Causation, for a relatively brief discussion).

How important was the choice of implicit bias model? Choice of implicit bias model clearly had some influence on how semicompatibilism responded, that said, I believe my critique of semicompatibilism is valuable and sound, based on a broadly accepted, nuanced and defensible understanding of implicit bias. From the complexity and disagreement within the free will/responsibility *and* implicit bias debates I have chosen particular positions and presented a critique of semicompatibilism in the context of a certain understanding of implicit bias; future research could, (and I would argue, should), look at freewill/responsibility and implicit bias across a broad range of free will and implicit bias perspectives.

The structure of this Thesis is straightforward, from a clear position on semicompatibilism (Part I) and implicit bias (Part II), semicompatibilism was critically examined with view to showing problems with semicompatibilism that arise when considering implicit bias influenced behaviour (Part III). After careful examination it was found that John Martin Fischer's semicompatibilism responded to the challenge of implicit bias; it was found that related behaviour *is* subject to guidance control, a conclusion in harmony with the models of implicit bias supported by much credible theory and practice. After showing this fundamental point, the focus of attention widened to examine an important defence of compatibilism from perhaps its biggest general challenge, the moral luck problem. The luck problem for compatibilism was chosen not only because it is one of the most substantial problems faced by compatibilists, but importantly, implicit bias appeared to be a paradigm example of a source of behaviour formed and often continually reinforced by factors that are subject to luck. Having looked specifically at the direct challenge of implicit bias for the semicompatibilist position itself, it was appropriate to investigate a defence of semicompatibilism from its biggest threat in the context of implicit bias. I concluded that Cyr's defence of semicompatibilism from the luck problem was *not* affected considering a model of implicit bias when implicit attitudes were included within an agent's deliberative standpoint. Levy's alternative characterisation of implicit bias takes such attitudes outside of an agent's deliberative standpoint, unavailable to Cyr's defence, hence subject to constitutional luck. Importantly then, compatibilism remained threatened by the luck problem in the context of agent implicit bias characterised by Levy and Cyr's defence. This is important because Cyr's defence of semicompatibilism from perhaps its most significant threat is at risk *if* Levy's understanding of implicit bias is correct.

The problem that luck brings to semicompatibilism is indicative of a more general problem described in Chapter 2, (page 50), the problem of manipulation. Helen Steward, quoted by John Martin Fischer *My Way and Life's Highway: Replies to Steward, Smilansky, and Perry* (2008b), crystallises her scepticism of semicompatibilist accounts because of the problem of manipulation:

> It appears to me, moreover, that this same basic difficulty (for a mechanism to be 'the agent's own') is going to infect any view that fails to assign actions the sort of metaphysically exceedingly distinctive nature I have been insisting they must have (See Appendix B). Any view which descends from the level of agents to the level of such things as mechanisms, processes and events is going to face the problem that any mechanism, process or event which occurs inside an agent can be set in train by someone, or something, which is not the agent. Only if one accepts that an action is essentially the exercise of a power by the agent whose action it is can this difficulty possibly be avoided. (Fischer 2008b: 173)

The problem, as previously described, is created by the possibility that some form of agent manipulation 'or event … set in train by someone, or something, which is not the agent' could reproduce and maintain the contents of the agent's mechanism, the actual mechanism that issues in behaviour. Under these circumstances the agent is surely not acting freely and responsibly. (It is interesting to reflect on the role of Frankfurt's controller in this context). This can lead to scepticism concerning a central plank of semicompatibilism; the very possibility of the actual mechanism that issues in behaviour being *truly* 'the agent's own'. John Martin Fischer would not accept that under conditions of manipulation the issuing mechanism *was* the agent's own and would argue that an agent was thereby not responsible for manipulated actions. Ownership and responsibility for the issuing mechanism for John Martin Fischer is the outcome of a detailed and subtle process that occurs over time, and yet, Steward's argument continues to be troubling.

I mention this problem again within the Conclusion of this Thesis because manipulation, luck and the implicit influence of bias on behaviour, are all threats, (they may all be used as a basis of claims and arguments that threaten), agent ownership of the actual sequence mechanism and present *the* most significant difficulties for semicompatibilism.[218] I have looked in detail at implicit bias and considered relatively briefly the luck problem and the very notion of agency. John Martin Fischer obviously takes the manipulation threat seriously; 'manipulation' is mentioned eighty-five times in *Responsibility and Control: A Theory of Moral Responsibility* (1998) and Papers have been written specifically addressing manipulation, for example, *Responsibility and Manipulation* (2004). In reply to Steward, John Martin Fischer maintains that moral responsibility is a matter of how the actual sequence proceeds and

> it does not follow that such responsibility is expunged by facts about the distal, (situated away from the point of attachment or origin or a central point), features of the actual sequence, such as whether it is set in motion by an agent with certain intentions. So, for example, an agent can exercise guidance control along a certain sequence; that is, the sequence can contain behavior that issues from his own, appropriately reasons-responsive mechanism. Now whether this sequence was set in motion millions of years ago (or thousands or hundreds...) by an individual with the intention that it proceed just as it does seems to me to be entirely irrelevant to the agent's moral responsibility. (my addition Fischer 2008b: 174)

John Martin Fischer robustly defends his position, but based on this response, the discord between, 'the sequence can contain behavior that issues from his *own*, appropriately reasons-responsive mechanism' and 'set in motion millions of years ago … by an individual with the intention that it proceed just as it does', still appears problematic. However, for John Martin Fischer, 'his own, appropriately reasons-responsive mechanism', the detailed and carefully considered requirement of guidance control, *excludes* by definition examples of manipulation because in such cases the issuing

[218] Some would say Manipulation *and* the Consequence Argument are the biggest challenges to Compatibilism (Kapitan 2000). John Martin Fischer's semicompatibilism does not have to comply with the principle of alternate possibilities as a condition of (guidance) control and responsibility.

mechanism is *not* the agent's *own*. Introducing agent causation, (See Appendix B), into the semicompatibilist model could resolve the problem, but this would be too restrictive for John Martin Fischer's project by committing to a particular metaphysical position.

In Appendix A I look at the nature of mental representations responsible for biased behaviour and describe a position not built on propositional attitudes nor mere associations, (the two basic accounts of implicit bias, although the description 'basic account' is rarely appropriate), rather, based on the concept of mental imagery. Appendix B outlines some important considerations concerning agency. The possibility and nature of human agency is of vital concern within the free will debate. Much consideration is given within the free will debate to the apparent problem of agency in a world where determinism is true. While agency per se is not the principal enterprise of this Thesis, it is important to give via an Appendix an outline of some key issues, given their shared nature with the free will debate generally and with semicompatibilism in particular. I outline the standard conception of agency and standard theory of action, then briefly outline three metaphysical frameworks of agency and look at compatibilism, agency and emergentism. The idea of agent causation is found to be very plausible as an emergent human (and some nonhuman animal) property within an incompatibilist/libertarian model of free will.

For information Appendix C provides a plot summary of Sophocles *Oedipus Rex* c429 BCE discussed in Part I. An illustration of human freedom restricted or eliminated by factors beyond an agent's control yet retaining the possibility of meaningful responsible choices. A recurring theme, which may be seen ultimately in semicompatibilism. I believe it was important to locate semicompatibilism within an historical process rather than simply in isolation.

In final conclusion, from an historical perspective I have looked at some key issues concerning threats to human freedom. This concluded with a detailed description of semicompatibilism, the free will position chosen for investigation considering the phenomena of implicit bias. While the nature of implicit bias is controversial, a characterisation of implicit bias was developed based on what I believe to be substantial and sound research. The semicompatibilist position on free will and responsibility, developed by John Martin Fischer, was found not to be threatened by the challenge of implicit bias. Implicit bias related behaviour was shown to be subject to guidance control and so agent responsibility in accord with the models of implicit bias developed in Part II: No problems with semicompatibilism/guidance control were found when considering implicit bias influenced behaviour.

Does Implicit Bias Threaten the Semicompatibilist Position on *Free Will* and Responsibility? Guidance control bypasses various metaphysical issues concerning the compatibility of moral *responsibility* and causal determinism; '… causal determinism is compatible with moral responsibility, … moral responsibility does not require genuine metaphysical access to alternative possibilities …' (Fischer 2015b: 203). Implicit bias has no impact on guidance control's separation from such metaphysical issues; there is nothing intrinsic to implicit bias that causes a problem for semicompatibilism's essentially agnostic position on free will, expressed in terms of metaphysical access to alternative possibilities, such concerns may be respectfully put to one side. Also, there is nothing intrinsic to implicit bias that causes a problem for semicompatibilism and responsibility.

It was found that Cyr's defence of semicompatibilism from the luck problem was not affected when considering the first model of implicit bias when implicit attitudes are included within an agent's deliberative standpoint. However, compatibilism is still threatened by the luck problem in the context of implicit bias characterised by Levy.

A constant theme throughout this Thesis has been awareness. For example, awareness of the wrongness of actions and awareness of the presence and/or issuing behaviour of implicit bias. As noted from the beginning of this Thesis, much relies on our understanding of awareness, therefore it is worth confirming again, within this final concluding section, that a position has been taken within this Thesis, based on good theoretical and empirical support, that the notion of implicit bias as a totally unconscious phenomenon is not supported by available empirical evidence. That said, under stress an agent's ability to deliberate is reduced with responsibility likewise lessened. The final words on this matter are from Madva, who argues persuasively and at length in Section 2 of *Implicit Bias, Moods, and Moral Responsibility* (2018) that empirical evidence supports agent awareness of their implicit biases, concluding emphatically that 'While many empirical questions remain unanswered, it seems clear that we cannot cast implicit biases into what popular authors … call "the locked door of the unconscious" '(2018: 60).

Future research could consider other free will/responsibility/agency and implicit bias positions and perspectives, aiming at a broader appraisal of the impact of the phenomenon of implicit bias on free will and responsibility. Given current (January 2021) racial, political and social upheaval, the nature of implicit bias is under ever increasing scrutiny. Future research, taking account of new knowledge such ongoing enquiry creates, would continue to bring together in important ways the long-established theme of free will/responsibility and implicit bias scholarship.[219]

I believe the research, arguments and conclusions presented above make an original contribution to knowledge and the free will debate.

~~~~~~

[219] While discussing J. S. Mill *On Liberty*, Oskari Kuusela describes the vital role *of* philosophy in achieving freedom, that 'to *engage philosophically* with obstacles to human freedom may be a liberating and emancipatory experience: … more generally, we might sometimes not even be properly aware of the obstacles to our freedom; the obstacles may be modes of being and thinking we have adopted unconsciously and quite unnoticed. However, in so far as philosophy is capable of drawing our attention to such things and can help us find alternative ways of thinking and acting, it can be comprehended as a liberating and emancipatory practice … . Thus conceived, philosophy is something we may take up and use to transform ourselves and to *achieve* freedom' (added emphasis 2011: 38).
~~~~~~

Bibliography

Agnew, Christopher, Donal Carlston, William Graziano, and Janice Kelly (eds.). 2010. *Then a Miracle Occurs: Focusing on Behaviour in Social Psychological Theory and Research* (Oxford: Oxford Scholarship Online) <https://doi.org/10.1093/acprof:oso/9780195377798.001.0001> [accessed 19 November 2020]

Andre, Judith. 1983. 'Nagel, Williams, and Moral Luck', *Analysis*, 43.4: 202–7

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

St. Aquinas. 1947. *The Summa Theologica, First Part (Q1-119) [1265 - 1274 C.E.]* <https://ccel.org/a/St. Aquinas/summa/FP/FP083.html#FPQ83OUTP1> [accessed 5 July 2017]

———. 1955. *Summa Contra Gentiles, Book One: God, Chapters 1 - 102 [1259–1265 C.E.]* <http://dhspriory.org/thomas/ContraGentiles1.htm#67> [accessed 11 July 2017]

Athanassoulis, Nafsika. 2017. 'Virtue Ethics', *Internet Encyclopedia of Philosophy* <http://www.iep.utm.edu/virtue/> [accessed 23 June 2017]

Augustine. 1871. 'Saint St. Augustine Complete Works [354–430 C.E.]', *Internet Archive 2011*, ed. by Marcus Dods (Edinburgh: T and T Clark) <https://archive.org/stream/AugustineTheWorksNewTranslation/Augustine-The Works, New Translation)#page/n213/mode/2up> [accessed 5 July 2017]

Ayala-López, Saray, and Erin Beeghly. 2020. 'Explaining Injustice: Structural Analysis, Bias, and Individuals', *An Introduction to Implicit Bias*, Kindle (Oxford: Routledge), pp. 211–32

Bagnoli, Carla. 2018. 'Claiming Responsibility for Action Under Duress', *Ethical Theory and Moral Practice*: 1–18 <https://doi.org/10.1007/s10677-018-9904-8>

Banaji, Mahzarin R., and Anthony G. Greenwald. 2013. *Blindspot: Hidden Biases of Good People* (New York: Delacorte Press)

Bargh, J.A. 1999. 'The Cognitive Monster: The Case Against Controllability of Automatic Stereotype Effects', *Dual Process Theories in Social Psychology* (New York:

Guilford Press), pp. 361–82

Bargh, John, and Ezequiel Morsella. 2010. 'Unconscious Behavioral Guidance Systems', *Then A Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research*, ed. by Christopher Agnew, Donal Carlston, William Graziano, and Janice Kelly (Oxford Scholarship Online), pp. 1–36 <https://doi.org/10.1093/acprof:oso/9780195377798.001.0001>

Beeghly, Erin. 2020. 'Bias and Knowledge. Two Metaphors', *An Introduction to Implicit Bias*, ed. by Beeghly, Erin, and Alex Madva, Kindle (Oxford: Routledge) <https://doi.org/10.4324/9781315107615>

Beeghly, Erin, and Alex Madva (eds.). 2020. *An Introduction to Implicit Bias*, Kindle (Oxford: Routledge) <https://doi.org/10.4324/9781315107615>

Berger, Jacob. 2018. 'Implicit Attitudes and Awareness', *Synthese*, pp. 1–22 <https://link.springer.com/journal/11229> [accessed 3 July 2019]

Berman, Sophie. 2004. 'Human Free Will in Anselm and Descartes', *The Saint Anselm Journal*, 2.1: 1–9

Blair, Irene V. 2002. 'The Malleability of Automatic Stereotypes and Prejudice', *Personality and Social Psychology Review*, 6.3: 242–61

Blanton, H, J Jaccard, J Klick, B Mellers, G Mitchell, and others. 2009. 'Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT', *Journal of Applied Psychology*, 94.3: 567–82

Boswell, James. 1998. *The Life of Samuel Johnson* (The Folio Society) Vol. 2: Entry for 16th October 1769.

Brandenburg, Daphne. 2016. 'Implicit Attitudes and the Social Capacity for Free Will', *Philosophical Psychology*, 29.8: 1215–28 <http://dx.doi.org/10.1080/09515089.2016.1235263>

Bratman, Michael. 2000. 'Review: Fischer and Ravizza on Moral Responsibility and History', *Philosophy and Phenomenological Research*, 61.2: 453–58 <https://www.jstor.org/stable/2653662>

Broad, C D. 1925. *The Mind and Its Place in Nature* (London: Kegan Paul, Trench. Trubner & Co., Ltd. (Digitized by the Internet Archive in 2010)) <http://www.archive.org/details/minditsplaceinnaOObroa>

Brockbank, Brien. 2019. 'Descartes and Scholasticism: An Analysis', *Aporia*, 29.1

Brownstein, Michael. 2016a. 'Attributionism and Moral Responsibility for Implicit Bias', *The Review of Philosophy and Psychology*, 7.4: 765–86 Springer Link https://link.springer.com/article/10.1007/s13164-015-0287-7 [accessed 1 January 2021]

———. 2016b. 'Implicit Bias', *Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta <https://plato.stanford.edu/archives/win2016/entries/implicit-bias/> [accessed 19 November 2020]

———. 2018. *The Implicit Mind: Cognitive Architecture, the Self, and Ethics* (Oxford: Oxford Scholarship Online) <https://doi.org/10.1093/oso/9780190633721.001.0001> [accessed 19 November 2020]

Brownstein, Michael, Alex Madva, and Bertram Gawronski. 2019. 'Understanding Implicit Bias: Putting the Criticisism into Perpective', *Pacific Philosophical Quarterly* (forthcoming)

Brownstein, Michael, and Jennifer Saul (eds.). 2016a. *Implicit Bias and Philosophy Volume 1, Metaphysics and Epistemology*, First edn. (Oxford University Press)

——— (eds.). 2016b. *Implicit Bias and Philosophy Volume 2, Moral Responsibility Structural Injustice and Ethics*, First edn. (Oxford: Oxford University Press)

Buckwalter, Wesley. 2018. 'Implicit Attitudes and the Ability Argument', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, pp. 1–30 <https://doi.org/10.1007/s11098-018-1159-7>

Capes, Justin. 2014. 'The Flicker of Freedom: A Reply to Stump', *The Journal of Ethics*, 18.4 427–35 <https://www.jstor.org/stable/43895888> [accessed 19 November 2020]

Carlson, Erik. 2003. 'Counterexamples to Principle Beta: A Response to Crisp and Warfield', *Philosophy and Phenomenological Research*, 66.3: 730–37 <https://www.jstor.org/stable/20140570> [accessed 19 November 2020]

Carlsson, Rickard, and Jens Agerstrom. 2015. *A Closer Look at the Discrimination Outcomes in the IAT Literature* (391 82 Kalmar Sweden)

Cartwright, Mark. 2013. 'Delphi', *Ancient History Encyclopedia* <https://www.ancient.eu/delphi/> [accessed 13 June 2017]

Caruso, Gregg D. 2015. 'Précis of Neil Levy's Consciousness and Moral Responsibility', *Journal of Consciousness Studies*, 22.7-8: 7-15

———. 2016. *Consciousness, Free Will, and Moral Responsibility* <https://ssrn.com/abstract=2888513> (Forthcoming in *Routledge Handbook of Consciousness*, ed. Rocco J. Gennaro [accessed 19 November 2020]

Clark, Andy. 2006. *Soft Selves and Ecological Control*, (Edinburgh: Philosophy Research Publications) <https://www.era.lib.ed.ac.uk/handle/1842/1446> [accessed 19 November 2020]

Clarke, Randolph. 2005. 'On an Argument for the Impossibility of Moral Responsibility', *Mid West Studies in Philosophy*, XXIX <https://doi.org/10.1111/j.1475-4975.2005.00103.x>

Clarke, Randolph, and Justin Capes. 2017. 'Incompatibilist (Nondeterministic) Theories of Free Will', *The Stanford Encyclopedia of Philosophy* ed. by Edward N Zalta <https://plato.stanford.edu/archives/spr2017/entries/incompatibilism-theories/> [accessed 4 October 2020]

Coates, D. Justin, and Philip Swenson. 2013. 'Reasons-Responsiveness and Degrees of Responsibility', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 165: 629–45 <https://doi.org/10.1007/s11098-012-9969-5>

Collins, Brian. 2013. 'Adding Substance to the Debate: Descartes on Freedom of the Will', *Essays in Philosophy*, 14.2: 218–38 <https://doi.org/10.7710/1526-0569.1473>

Collinson, Diané, and Kathryn Plant. 2006. *Fifty Major Philosophers* (Oxford: Routledge)

Cottingham, John (ed.). 1992. *The Cambridge Companion to Descartes* (New York: Cambridge University Press)

Cox, Gary. 2014. *Sartre: Consciousness, Freedom, Bad Faith*, Kindle edn.

Cudworth, Ralph. 1996. *A Treatise Concerning Eternal and Immutable Morality* , ed. by Sarah Hutton (Cambridge UK: Cambridge University Press)

Cyr, Taylor W. 2018. *Semicompatibilist Options: Essays in Defense of an Actual-Sequence Approach to Freedom and Responsibility* (University of California Riverside Doctoral Disertation) <https://escholarship.org/uc/item/9h245183> [accessed 4 October 2020]

Cyr, Taylor W. 2019. 'Moral Responsibility, Luck, and Compatibilism', *Erkenntnis*, 84.1: 193–214 <https://doi.org/10.1007/s10670-017-9954-7>

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

<http://www.oxfordscholarship.com/view/10.1093/0199246270.001.0001/acprof-9780199246274>

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

———. 2003. *Freedom Evolves* (London: Penguin Books)

Dennett, Daniel, and Gregg Caruso. 2021. *Just Deserts* (Cambridge: Polity Press)

Descartes, René. 1911. 'Meditations On First Philosophy', *The Philosophical Works of Descartes* (Cambridge University Press), pp. 1–33

———. 2012. 'Part 1: The Principles of Human Knowledge', *Principles of Philosophy [1644]*, Amended edn. <http://www.earlymoderntexts.com/assets/pdfs/descartes1644part1.pdf> [accessed 12 July 2017]

Deutsch, Roland, and Fritz Strack. 2010. 'Building Blocks of Social Behaviour', *Handbook of Implicit Social Cognition*, ed. by Bertram Gawronski and B. Keith Payne (New York: The Guilford Press), pp. 62–79

Devine, Patricia. 1989. 'Stereotypes and Prejudice: Their Automatic and Controlled Components', *Journal of Personality and Social Psychology*, 56.1: 5–18

Dilman, Ilham. 1999. *Free Will (An Historical and Philosophical Introduction)* (London: Routledge)

Doyle, Bob. 2017. 'Incompatibilism', *Information Philosopher* <http://www.informationphilosopher.com/freedom/incompatibilism.html> [accessed 12 September 2017]

Eberhardt, Jennifer. 2019. *Biased* (London: Penguin Random House)

Evans, Jonathan, and Keith Stanovich. 2013. 'Dual-Process Theories of Higher Cognition: Advancing the Debate', *Perspectives on Psychological Science*, 8.3: 223–41 <https://doi.org/10.1177/1745691612460685>

Evans, Jonathan. 2008. 'Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition', *The Annual Review of Psychology*, 59: 255–78 <https://doi.org/10.1146/annurev.psych.59.103006.093629>

Evans, Jonathan, and Keith Frankish. 2012. 'The Duality of Mind: An Historical Perspective', *In Two Minds: Dual Processes and Beyond* (Oxford Scholarship Online) <https://doi.org/10.1093/acprof:oso/9780199230167.001.0001>

Fanon, Frantz. 2007. *Black Skin, White Masks*, First edn. (New York: Grove Press / Grove Atlantic)

Fazio, Russell. 1990. 'Multiple Processes by Which Attitudes Guide Behavior: The MODE Model as an Integrative Framework', *Advances in Experimental Social Psychology*, 23: 75–109

Feinberg, Joel. 1989. 'Failures of Consent: Defective Belief', *The Moral Limits of the Criminal Law Volume 3: Harm to Self* (Oxford: Oxford Scholarship Online 2003) <https://doi.org/10.1093/0195059239.001.0001> [accessed 19 November 2020]

Fiedler, Klaus, and Momme Sydow. 2015. 'Heuristics and Biases: Beyond Tversky and Kahneman's (1974) Judgment under Uncertainty', *Cognitive Psychology: Revisiting the Classical Studies*, ed. by Michael W Eysenck and David Groome (London: Sage), pp. 146–61

Fischer, John Martin. 2003. 'Frankfurt-Style Compatibilism', *Free Will (Oxford Readings in Philosophy)*, Second edn., ed. by Gary Watson (Oxford: Oxford University Press), pp. 190–211

———. 2004. 'Responsibility and Manipulation', *The Journal of Ethics*, 8.2: 145–77

———. 2006a. *My Way: Essays on Moral Responsibility* (Oxford: Oxford University Press)

———. 2006b. 'Responsiveness and Moral Responsibility', *My Way Essays on Moral Responsibility* (New York: Oxford University Press, Inc.), pp. 63–83

———. 2008a. 'Freedom, Foreknowledge, and Frankfurt: A Reply to Vihvelin', *Canadian Journal of Philosophy*, 38.3: 327–42

———. 2008b. 'My Way and Life's Highway: Replies to Steward, Smilansky, and Perry', *The Journal of Ethics*, 12.2: 167–89 <https://doi.org/10.1007/s10892-008-9029-8>

———. 2010a. 'Précis of My Way: Essays on Moral Responsibility (2006)', *Philosophy and Phenomenological Research*, 80.1: 229–41 <https://www.jstor.org/stable/40380509>

———. 2010b. 'The Frankfurt Cases: The Moral of the Stories', *Philosophical Review*, 119.3: 315–36 <https://doi.org/10.1215/00318108-2010-002>

———. 2012. 'Semicompatibilism and Its Rivals', *Journal of Ethics*, 16: 117–43 <https://doi.org/10.1007/s 10892-0 12-9 123-9>

———. 2013. 'My Compatibilism', *The Philosophy of Free Will*, ed. by Paul Russell and Oisin Deery (Oxford: Oxford University Press), pp. 296–317

———. 2015a. 'Deep Control', *Deep Control: Essays on Free Will and Value* (Oxford: Oxford Scholarship Online) <https://doi.org/10.1093/acprof:osobl/9780199742981.001.0001> [accessed 19 November 2020]

———. 2015b. 'Guidance Control', *Deep Control: Essays on Free Will and Value* (Oxford: Oxford Scholarship Online) <https://doi.org/10.1093/acprof:osobl/9780199742981.001.0001> [accessed 19 November 2020]

———. 2015c. 'Sourcehood', *Deep Control: Essays on Free Will and Value* (Oxford: Oxford Scholarship Online) <https://oxford.universitypressscholarship.com/view/10.1093/acprof:osobl/9780199742981.001.0001/acprof-9780199742981-chapter-10> [accessed 27 November 2020]

———. 2016. 'How Do Manipulation Arguments Work?', *The Journal of Ethics*, 20.1–3: 47–67 <https://doi.org/10.1007/s10892-016-9225-x>

Fischer, John Martin, Robert Kane, Derk Pereboom, and Manuel Vargas. 2007. *Four Views on Free Will* (Oxford: Blackwell Publishing)

Fischer, John Martin, and Mark Ravizza. 1992. 'Responsibility, Freedom, and Reason (Review of Freedom Within Reason by Susan Wolf)', *Ethics*, 102.2: 368–89 <http://www.jstor.org/stable/2381611> [accessed 19 November 2020]

Fischer, John Martin, and Mark Ravizza (eds.). 1993. *Perspectives on Moral Responsibility* (New York: Cornell University Press)

———. 1998. *Responsibility and Control: A Theory of Moral Responsibility (Cambridge Studies in Philosophy and Law)*, Kindle edn. (Cambridge: Cambridge University Press)

———. 2000. 'Précis of Responsibility and Control: A Theory of Moral Responsibility (1998 Edn.)', *Philosophy and Phenomenological Research*, 61.2: 441–45 <https://www.jstor.org/stable/2653660> [accessed 19 November 2020]

Fischer, John Martin, and Neal Tognazzini. 2009. 'The Truth about Tracing', *Noûs*, 43.3: 531–56 <https://doi.org/10.1111/j.1468-0068.2009.00717.x>

Flanagan, Owen, and Gregg D. Caruso. 2018. ' Third Wave Existentialism', *Neuroexitentialism*, ed. by Gregg D Caruso and Owen Flanagan (New York USA: Oxford University Press)

Forscher et al., Patrick. 2019. *A Meta-Analysis of Procedures to Change Implicit Measures*, (University of Arkansas: Fayetteville Psychological Science Faculty Publications and Presentations) <https://devinelab.psych.wisc.edu/wp content/uploads/sites/1383/2020/04/A-Meta-Analysis-of-Procedures-to-Change-Implicit-Measures-1.pdf>

Foucault and Identity. 2020. *Changing Minds* <http://changingminds.org/explanations/identity/foucault_identity.htm> [accessed 25 September 2020]

Franchi, Leo. 2020. *Sartre and Freedom* <files.lfranchi.com/papers/sartre.and.freedom.pdf · PDF file> [accessed 22 September 2020]

Frankfurt, Harry. 1969. 'Alternate Possibilities and Moral Responsibility', *The Journal of Philosophy*, 66.23: 829–39

———. 1971. 'Freedom of the Will and the Concept of the Person', *The Journal of Philosophy*, 68.1: 5–20

Frankish, Keith. 2012. 'Systems and Levels: Dual-System Theories and the Personal - Subpersonal Distinction', *In Two Minds: Dual Processes and Beyond*, ed. by Jonathan Evans and Keith Frankish (University Press Scholarship Online) <https://doi.org/10.1093/acprof:oso/9780199230167.001.0001> [accessed 19 November 2020]

———. 2016. 'Playing Double - Implicit Bias, Dual Levels, and Self-Control', *Implicit Bias and Philosophy Volume 1*, First edn., ed. by Michael Brownstein and Jennifer Saul (Oxford: Oxford University Press), pp. 23–46

Franklin, Christopher Evan. 2013. 'Causes, Laws, and Free Will: Why Determinism Doesn't Matter (Review)', *Philosophical Reviews (University of Notre Dame)* <https://ndpr.nd.edu/news/causes-laws-and-free-will-why-determinism-doesn-t-matter/> [accessed 17 August 2018]

Freud, Sigmund. 1915. 'The Unconscious (1915)', *General Psychological Theory (The Collected Papers of Sigmund Freud, 1963)*, ed. by Philip Rieff (New York: Collier Books: Macmillan Publishing Company), pp. 116–50

Fridland, Ellen. 2017. 'Automatically Minded', *Synthese*, 194.11: 4337–63 <https://doi.org/10.1007/s11229-014-0617-9>

Gaertner, Samuel L., and J.P. McLaughlin. 1983. 'Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics', *Social Psychology Quarterly*, 46.1: 23–30

Gawronski, Bertram. 2019. 'Six Lessons for a Cogent Science of Implicit Bias and Its Criticism', *Perspectives on Psychological Science*, 14.4: 574–95 <https://doi.org/DOI: 10.1177/1745691619826015>

Gawronski, Bertram, Eva Walther, and Hartmut Blank. 2005. 'Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information', *Journal of Experimental Social Psychology*, 41: 618–626

Gawronski, Bertram, and Galen Bodenhausen. 2006. 'Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change', *Psycological Bulletin*, 132: 692–731

Gawronski, Bertram, and Laura Creighton. 2013. 'Dual Process Theories', *The Oxford Handbook of Social Cognition*, ed. D Carlston (New York: Oxford University Press), pp. 282–312

Gawronski, Bertram, Wilhelm Hofmann, and Christopher Wilbur. 2006. 'Are "Implicit" Attitudes Unconscious?', *Consciousness and Cognition*, 15: 485–99

Gawronski, Bertram, and B. Keith Payne (eds.). 2010. *Handbook of Implicit Social Cognition* (New York: The Guilford Press)

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and Biases; The Psychology of Intuitive Judgement* (Cambridge: Cambridge University Press)

Goodwin, Michele. 2018. 'Revisiting Death: Implicit Bias and the Case of Jahi McMath', *Hastings Center Report*, 48.S4: 77–80

Greenwald, Anthony G., Mahzarin R. Banaji, Laurie A. Rudman, Shelly D. Farnham, Brian A. Nosek, and others. 2002. 'A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept', *Psychological Review*, 109.1: 3–25 <https://doi.org/10.1037/0033-295X.109.1.3>

Greenwald, Anthony G., and Linda Krieger. 2006. 'Implicit Bias: Scientific Foundations', *California Law Review*, 94.4: 945–67

Greenwald, Anthony G, Debbie E McGhee, and Jordan L K Schwartz. 1998. 'Measuring Individual Differences in Implicit Cognition: The Implicit Association Test', *Journal of Personality and Social Psychology*, 74.6: 1464–80 <http://faculty.fortlewis.edu/burke_b/Senior/BLINK replication/IAT.pdf> [accessed 7 November 2017]

Gregg, Aiden, Mahzarin R. Banaji, and Beate Seibt. 2006. 'Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences', *Journal of Personality and Social Psychology*, 90.1: 1–20 <https://doi.org/10.1037/0022-3514.90.1.1>

Grim, Patrick. 2007. 'Free Will in Context: A Contemporary Philosophical Perspective', *Behavioral Sciences and the Law*, 25: 183–201

Hahmias, Eddy. 2001. *Free Will and the Knowledge Condition* (Durham, North Carolina USA: Duke University) <https://www.academia.edu/2734134/Free_will_and_the_knowledge_condition> [accessed 30 October 2019]

———. 2008. *Oxford Handbook on Philosophy of Psychology* (Unpublished), ed. by Jesse Prinz (Available on PhilArchive: https://philarchive.org/archive/NAHTPO-4)

Hannikainen, Ivar R., Edouard Machery, David Rose, Stephen Stich, Christopher Y. Olivola, and others. 2019. 'For Whom Does Determinism Undermine Moral Responsibility? Surveying the Conditions for Free Will Across Cultures', *Frontiers in Psychology*, 10 <https://doi.org/10.3389/fpsyg.2019.02428>

Hartman, Robert. 2018. 'Constitutive Moral Luck and Strawson's Argument for the Impossibility of Moral Responsibility', *Journal of the American Philosophical Association,* 4.2: 165–183 (also available from <www.robertjhartman.com>)

Heaney, Seamus. 2000. *Beowulf* (London: Faber and Faber)

Heider, Fritz. 1958. *The Psychology of Interpersonal Relations* (New York: Wiley)

Hemati, Russell Danesh. 2010. *Augustine's Solution to the Problem of Theological Fatalism* (Baylor University)

Hempel, Carl, and Paul Oppenheim. 1948. 'Studies in the Logic of Explanation', *Philosophy of Science*, 15.2: 135–75 <http://links.jstor.org/sici?sici=0031-8248%28194804%2915%3A2%3C135%3ASITLOE%3E2.0.CO%3B2-E>

Hintikka, Jaakko. 1991. *Knowledge and the Known* (The Netherlands: Kluwer Academic Publishers)

Holliday, Wesley. 2012. 'Freedom and the Fixity of the Past', *Philosophical Review*, 121.2: 179–207 <https://doi.org/10.1215/00318108-1539080>

Holroyd, Jules. 2010. 'Punishment and Justice', *Social Theory and Practice*, 36.1: 78–111 <www.jstor.org/stable/23558593> [accessed 1 March 2021]

———. 2012. 'Responsibility for Implicit Bias', *Journal of Social Philosophy*, 43.3: 274–306 <https://doi.org/10.1111/j.1467-9833.2012.01565.x>

Holroyd, Jules, and Daniel Kelly. 2016. 'Implicit Bias, Character, and Control', *From Personality to Virtue: Essays on the Philosophy of Character*, First edn., ed. by Alberto Masala and Jonathan Webber (Oxford: Oxford University Press (Oxford on Line)), pp. 106–33 <https://doi.org/10.1093/acprof:oso/9780198746812.003.0006> [accessed 19 November 2020]

Holroyd, Jules, Robin Scaife, and Tom Stafford. 2017. 'What Is Implicit Bias?', *Philosophy Compass*, 12.10: 1–18 <https://doi.org/10.1111/phc3.12437> [accessed 14 July 2019]

Holroyd, Jules, and Joseph Sweetman. 2016. 'The Heterogeneity of Implicit Bias', *Implicit Bias and Philosophy*, First edn., ed. by Michael Brownstein and Jennifer Saul (Oxford: Oxford University Press), pp. 80–103

Hume, David. 2004. 'An Enquiry Concerning Human Understanding, Liberty and Necessity Section 8, Parts 1 and 2', *Early Modern Texts* <http://www.earlymoderntexts.com/assets/pdfs/hume1748_2.pdf> [accessed 20 July 2017]

Hutcheson, Francis. 2017. *An Inquiry Into The Original Of Our Ideas of Beauty And Virtue* (California, USA: CreateSpace Independent Publishing Platform)

van Inwagen, Peter. 2003. 'An Argument for Incompatibilism', *Free Will (Oxford Readings in Philosophy)*, Second edn., ed. by Gary Watson (Oxford: Oxford University Press), pp. 38–57

Irwin, Terence. 1977. *Plato's Moral Theory, The Early and Middle Dialogues* (Oxford: Oxford University Press)

Jarvilehto, L. 2015. 'The Nature of Intuitive Thought', *The Nature and Freedom of Intuitive Thought and Decision Making*, First edn. (Springer International Publishing), pp. 23–54 <https://doi.org/10.1007/978-3-319-18176-9>

Johnson, Gabbrielle M. 2020. 'The Psychology of Bias: From Data to Theory', *An Introduction to Implicit Bias*, Kindle edn., ed. by Erin Beeghly and Alex Madva (Oxford: Routledge), pp. 20–41

Johnson, Robert, and Adam Cureton. 2016. 'Kant's Moral Philosophy', *The Stanford Encyclopedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/kant-moral/#CatHypImp> [accessed 26 September 2017]

Kahneman, Daniel. 2003. 'A Perspective on Judgment and Choice', *American Psychologist*, 58.9: 697–720 <https://doi.org/10.1037/0003-066X.58.9.697>

Kane, Robert (ed.). 2012. 'The Oxford Handbook of Free Will', *Oxford Handbooks Online* (Oxford: Oxford University Press) <https://doi.org/10.1093/oxfordhb/9780195399691.003.0027>

———. 2015. *Introduction (Oxford Handbook of Free Will)*, (Oxford: Oxford Handbooks Online) <https://doi.org/10.1093/oxfordhb/9780199552153.013.0032> [accessed 5 October 2020]

Kang, Jerry. 2005. 'Trojan Horses of Race', *Harvard Law Review*, 118.5: 1489–1593

Kant, Immanuel. 1889. *Kant's Critique of Practical Reason and Other Works*, (London: Longmans, Green and Co. Paternoster-Row)

Kaufman, Daniel. 2020. 'Value and Objectivity', *The Electric Agora* <https://theelectricagora.com/> [accessed 25 September 2020]

Kim, Jaegwon. 2007. 'Mental Causation and Consciousness: Our Two Mind-Body Problems', *Physicalism, or Something Near Enough* (Princeton NJ: Princeton University Press), pp. 7–31 <https://doi.org/10.1515/9781400840847.7>

———. 2009. 'Supervenience, Emergence, Realization, Reduction', *The Oxford Handbook of Metaphysics*, Online edn., ed. by Michael Loux and Dean Zimmerman (Oxford: Oxford Handbooks Online), pp. 1–31 <https://doi.org/10.1093/oxfordhb/9780199284221.003.0019> [accessed 5 October 2020]

King, M., and P. Carruthers. 2012. 'Moral Responsibility and Consciousness', *The Journal of Moral Philosophy*, 9: 200–228

Kuusela, Oskari. 2011. *Key Terms in Ethics* (London: Continuum International Publishing Group)

Lennon, Thomas M. 2013. 'Descartes's Supposed Libertarianism: Letter to Mesland or Memorandum Concerning Petau?', *Journal of the History of Philosophy*, 51.2: 223–48 <https://doi.org/10.1353/hph.2013.0026>

Leslie, Sarah-Jane. 2017. 'The Original Sin of Cognition: Fear, Prejudice, and Generalization', *The Journal of Philosophy*, CXIV.8: 393–421 <https://www.princeton.edu/~sjleslie/The original sin of cognition upd010518.pdf>

Levy, Neil. 2011a. 'An Account of Luck', *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford: Oxford Scholarship Online), pp. 12–40 <https://doi.org/10.1093/acprof:oso/9780199601387.001.0001> [accessed 5 October 2020]

———. 2011b. 'The Luck Problem for Compatibilists', *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford: Oxford Scholarship Online), pp. 85–109 <https://doi.org/10.1093/acprof:oso/9780199601387.001.0001> [accessed 5 October]

———. 2014a. 'Consciousness, Implicit Attitudes and Moral Responsibility', *Noûs*, 48.1: 21–40 <https://doi.org/10.1111/j.1468-0068.2011.00853.x>

———. 2014b. *Consciousness and Moral Responsibility* (Oxford: Oxford University Press) <https://doi.org/10.1093/acprof:oso/9780198704638.001.0001>

———. 2017. 'Implicit Bias and Moral Responsibility: Probing the Data', *Philosophy and Phenomenological Research*, 94.1: 1–20 <https://doi.org/10.1111/phpr.12352>

———. 2018. 'Michael Brownstein, The Implicit Mind: Cognitive Architecture, the Self, and Ethics, Oxford University - A Review by Neil Levy, Macquarie University/University of Oxford', *Notre Dame Philosophical Reviews 05.09.2018* <https://ndpr.nd.edu/news/the-implicit-mind-cognitive-architecture-the-self-and-ethics/>

Lewis, David. 1981. 'Are We Free to Break the Laws?', *Theoria*, 47.3: 113–21 <https://doi.org/10.1111/j.1755-2567.1981.tb00473.x>

Long, Todd R. 2004. 'Moderate Reasons-Responsiveness, Moral Responsibility, and Manipulation', *Freedom and Determinism*, ed. by Joseph Campbell, Michael O'Rourke, and David Shier (Massachusetts: MIT Press), pp. 151–72

Madva, Alex. 2018. 'Implicit Bias, Moods, and Moral Responsibility', *Pacific Philosophical*

Quarterly, 99.S1: 53–78

———. 2020. 'Implicit Bias', *Ethics in Practice: An Anthology*, Fifth edn., ed. by Hugh LaFollette (Chichester: John Wiley and Sons Inc), pp. 387–95

Magee, Bryan. 1978. *Men of Ideas* (London: BBC Books)

———. 1987. *The Great Philosophers* (London: BBC Books)

Mandelbaum. 2016. 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias', *Noûs*, 50.3: 629–58 <https://doi.org/10.1111/nous.12089> Note: Paper sourced from PhilPapers <https://philpapers.org/rec/MANAIA-5> In text Citation page numbers refer to PhilPapers sourced document. [accessed 5 October 2020]

———. 2017. 'Associationist Theories of Thought', *Stanford Encyclopaedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/associationist-thought/> [accessed 5 October 2020]

Mason, Elinor, and Alan T Wilson. 2017. 'Vice, Blameworthiness, and Cultural Ignorance', *Responsibility: The Epistemic Condition*, Online edn., ed. by Philip Robichaud and Jan Willem Wieland (Oxford: Oxford University Press (Oxford on Line)), pp. 83–98 <https://doi.org/10.1093/oso/9780198779667.003.0004>

McKenna, Michael, and D. Justin Coates. 2015a. 'Compatibilism', *Stanford Encyclopedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/compatibilism/> [accessed 2 September 2017]

———. 2015b. 'Compatibilsm: The State of the Art', *Stanford Encyclopedia of Philosophy* (Stanford University) <https://plato.stanford.edu/archives/spr2015/entries/compatibilism/supplement.html> [accessed 5 October 2020]

McLoughlin, Siobhán. 2012. *The Freedom of the Good: A Study of Plato's Ethical Conception of Freedom* (University of New Mexico UNM Digital Repository) <http://digitalrepository.unm.edu/phil_etds> [accessed 14 June 2017]

Mele, Alfred R. 2005. 'A Critique of Pereboom's "Four-Case Argument" for Incompatibilism', *Analysis*, 65.1: 75–80 <http://www.jstor.org/stable/3329340>

———. 2008. 'Manipulation, Compatibilism, and Moral Responsibility', *Journal of Ethics*, 12: 263–286

Mendelson, Michael. 1997. 'Saint Augustine', *Stanford Encyclopedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/augustine/#Wil> [accessed 28 June 2017]

Moors, Agnes, and Jan De Houwer. 2006. 'Automaticity: A Theoretical and Conceptual Analysis', *Psycological Bulletin*, 132.2: 297–326 <https://doi.org/10.1037/0033-2909.132.2.297>

Morris, William Edward, and Charlotte R. Brown. 2014. 'David Hume', *The Stanford Encyclopedia of Philosophy (Spring 2017)* (Stanford University) <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=hume> [accessed 18 July 2017]

Moskowitz, Gordon, and Li Peizhong. 2011. 'Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control', *Journal of Experimental Social Psychology*, 47.1: 103–16

Murray, Samuel. 2019. 'The Place of the Trace: Negligence and Responsibility', *The Review of Philosophy and Psychology* <https://www.academia.edu/39888214/The_Place_of_the_Trace_Negligence_and_Responsibility?auto=download>

Nadelhoffer, Thomas, Jason Shepard, Eddy Hahmiaas, Chandra Sripada, and Lisa Thomson-Ross. 2014. 'The Free Will Inventory: Measuring Beliefs about Agency', *Consciousness and Cognition*, 25: 27–41 <www.elsevier.com/locate/concog>

Nagel, Thomas. 1979. *Mortal Questions* (Cambridge: Cambridge University Press)

———. 2003. 'Freedom', *Free Will (Oxford Readings in Philosophy)*, Second edn., ed. by Gary Watson (Oxford: Oxford University Press), pp. 229–56

Nanay, Bence. 2020. 'Implicit Bias as Mental Imagery', *Journal of the American Philosophical Association (Forthcoming)*

Nelkin, Dana. 2019. 'Moral Luck', *Stanford Encyclopaedia of Philosophy* (Stanford University) <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=moral-luck> [accessed 25 September 2020]

Newberger Goldstein, Rebecca. 2015. 'What Plato Tells Us About Your Moral Choices', *World Economic Forum* <https://www.weforum.org/agenda/2015/01/what-plato-tells-us-about-your-moral-choices/> [accessed 20 September 2019]

Nozick, Robert. 1982. *Philosophical Explanations* (USA Cambridge, Massachusetts: Belknap Press of Harvard University)

O'Connor, Timothy. 2013. 'Agent-Causal Power', *The Philosophy of Free Will*, ed. by Paul Russell and Oisin Deery (Oxford: Oxford University Press), pp. 229–48

Palmer, David. 2014. 'Deterministic Frankfurt Cases', *Synthese*, 191.16: 3847–64 <https://link.springer.com/article/10.1007%2Fs11229-014-0500-8>

Parks, Tim, and Riccardo Manzotti. 2020. 'You Are the World', *Aeon* (From the Book 'Dialogues on Consciousness' (2020) by Riccardo Manzotti and Tim Parks (New York: OR Books) <https://aeon.co/essays/the-question-is-what-are-we-a-conversation-on-consciousness> [accessed 25 March 2020]

Payne, B. Keith, Heidi A. Vuletich, and Jazmin Brown-Iannuzzi. 2019. 'Historical Roots of Implicit Bias in Slavery', *Proceedings of the National Academy of Sciences of the USA*, ed. by Jennifer Richeson (Washington, DC 20001 USA: National Academy of Sciences of the USA), pp. 11693–98 <https://doi.org/10.1073/pnas.1818816116>

Peels, Rik. 2014. 'What Kind of Ignorance Excuses? Two Neglected Issues', *Philosophical Quarterly*, 64: 478–96 <https://doi.org/10.1093/pq/pqu013>

Pereboom, Derk. 2000. 'The Significance of Free Will', *Ethics*, 110.2: 426–30 <https://doi.org/10.1086/233282>

———. 2003. *Living without Free Will (Cambridge Studies in Philosophy)*, (Cambridge: Cambridge University Press)

———. 2016. *Free Will, Agency, and Meaning in Life*, Reprint edn. (Oxford: Oxford University Press)

Plato. 2004. *Gorgias (Penguin Classics)*, Revised edn., ed. by Chris Emlyn-Jones (London: Penguin Books)

———. 2019. *The Essential Plato Anthology (25 Works)*, Kindle edn., ed. and translated by Benjamin Jowett

Ragland, C P. 2013. 'Descartes on Degrees of Freedom: A Close Look at a Key Text', *Essays Philosophy*, 14: 239–68 <https://doi.org/10.7710/1526-0569.1474>

Ravizza, Mark. 1993. *Semi-Compatibilism and the Transfer of Non-Responsibility* <https://www.sheffield.ac.uk/polopoly_fs/1.101519!/file/Ravizza-on-nontransfer-of-MR.pdf> [accessed 10 February 2019] Print copy available

Philosophical Studies 1994 75: 61-93

Reber, Arthur. 1989. 'Implicit Learning and Tacit Knowledge', *Journal of Experimental Psychology*, 118.3: 219–35

Rettler, Lindsay, and Bradley Rettler. 2019. 'Epistemic Duty and Implicit Bias', *Epistemic Duties*, ed. by Kevin McCain and Scott Stapleford (London: Routledge)

Rohlf, Michael. 2016. 'Immanuel Kant', *The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)*, ed. by Edward N. Zalta (Stanford University) <https://plato.stanford.edu/archives/spr2016/entries/kant/> [accessed 20 November 2020]

Rudman, Laurie A. 2004. 'Social Justice in Our Minds, Homes, and Society: The Nature, Causes, and Consequences of Implicit Bias', *Social Justice Research* (Norwell, MA 02061 USA: Kluwer Academic Publishing), 17.2: 129–42

Rudy-Hiller, Fernando. 2018. 'The Epistemic Condition for Moral Responsibility', *The Stanford Encyclopedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/moral-responsibility-epistemic/> [accessed 27 August 2019]

Russell, Paul. 2011. 'Moral Sense and the Foundations of Responsibility', *The Oxford Handbook of Free Will: Second Edition*, ed. by Robert Kane (Oxford: Oxford University Press), pp. 199–220

———. 2013a. 'Free Will and Responsibility', *Five Books* <http://fivebooks.com/interview/paul-russell-on-free-will-and-responsibility/> [accessed 5 May 2017]

——— (ed.). 2013b. *The Philosophy of Free Will: Essential Readings from the Contemporary Debates* (Oxford University Press)

———. 2014. 'Hume on Free Will', *Stanford Encyclopedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/hume-freewill/> [accessed 20 July 2017]

———. 2015a. *Freewill Pessimism (New Orleans Workshop on Agency and Responsibility)* <https://www.academia.edu/17543865/Free_Will_Pessimism> [accessed 5 October 2020] Also available *Oxford Studies in Agency and Responsibility* 2017 ed. David Shoemaker, Volume 4, Chapter 5, DOI: 10.1093/oso/9780198805601.001.0001

———. 2015b. ' "Hume's Lengthy Digression": Free Will in the Treatise', *Hume's Treatise: A Critical Guide*, ed. by A Butler and D Ainslie (Cambridge: Cambridge University Press), pp. 230–51

Rychter, Pablo. 2017. 'Does Free Will Require Alternative Possibilities?', *Disputatio International Journal of Philosophy*, 9.45: 131–46 <https://doi.org/https://doi.org/10.1515/disp-2017-0001>

Salter, Phia S., Glenn Adams, and Michael J. Perez. 2018. 'Racism in the Structure of Everyday Worlds: A Cultural-Psychological Perspective', *Current Directions in Psychological Science*, 27.3: 150–55 <https://doi.org/10.1177/0963721417724239>

Sartorio, Carolina. 2014. 'Vihvelin on Frankfurt-Style Cases and the Actual- Sequence View', *Criminal Law and Philosophy* (2016 issue date) 10: 875–888 <https://link.springer.com/article/10.1007/s11572-014-9355-9

Sartre, Jean-Paul. 1961. *The Age of Reason* (Harmondsworth, England: Penguin Books)

———. 1963. *The Reprieve* (Harmondsworth, England: Penguin Books)

———. 1963. *Iron in the Soul* (Harmondsworth, England: Penguin Books)

———. 1976. *Anti-Semite and Jew* (New York: Schocken Books)

———. 1984. *Being and Nothingness* (New York: Washington Square Press)

———. 1992. *Notebooks for an Ethics* (Chicago: University of Chicago Press)

———. 2004. *Critique of Dialectical Reason Volume One* (London: Verso)

Saunders, John Turk. 1968. 'The Temptations of "Powerlessness" ', *American Philosophical Quarterly*, 5.2: 100–108 <https://www.jstor.org/stable/20009261>

Schlosser, Markus. 2019. 'Agency', *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)* (Stanford University) <https://plato.stanford.edu/entries/agency/> [accessed 5 October 2020]

Schmaltz, Tad. 2008. *Descartes on Causation* (New York: Oxford University Press Inc.)

Scruton, Roger. 2012. *Modern Philosophy, An Introduction and Survey* (London: Bloomsbury Reader), pp. 227

Sher, George. 2017. 'Blame and Moral Ignorance', *Responsibility: The Epistemic Condition*, Online edn., ed. by Philip Robichaud and Jan Willem Wieland (Oxford: Oxford University Press (Oxford on Line)), pp. 102–16 <https://doi.org/10.1093/oso/9780198779667.003.0005>

Smilansky, Saul. 2012. 'Free Will and Moral Responsibility: The Trap, the Appreciation of Agency, and the Bubble', *The Journal of Ethics*, 16.2: 211-239.

Smith, Eliot, and Jamie De Coster. 2000. 'Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems', *Personality and Social Psychology Review*, 4.2: 108–31 <https://doi.org/10.1207/S15327957PSPR0402_01>

Smith, Holly. 1983. 'Culpable Ignorance', *The Philosophical Review*, 92.4: 543–71

Sorensen, David. 2016. *The Unity of Higher Cognition: The Case against Dual Process Theory*, (U.S.A: Georgia State University) <https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1197&context=philosophy_theses>

Speak, Daniel. 2008. 'Guest Editor's Introduction: Leading the Way', *The Journal of Ethics*, 12.2: 123–28 <https://doi.org/10.1007/s10892-008-9026-y>

Staats, Cheryl, Capatosto, Kelly, Lena Tenney, and Sarah Mamo. 2017. *State of the Science: Implicit Bias Review 2017* (Columbus, Ohio: Kirwan Institute) <http://kirwaninstitute.osu.edu/2017-state-of-the-science-implicit-bias-review/> [accessed 3 September 2019]

Staats, Cheryl. 2013. *State of the Science: Implicit Bias Review 2013* (Columbus, Ohio: Kirwan Institute) <http://www.kirwaninstitute.osu.edu/reports/2013/03_2013_SOTS-Implicit_Bias.pdf> [accessed 3 September 2018]

Staats, Cheryl, Kelly Capatosto, Robin Wright, and Victoria Jacksom. 2016. *State of the Science: Implicit Bias Review 2016* (Columbus, Ohio: Kirwan Institute) <http://kirwaninstitute.osu.edu/my-product/2016-state-of-the-science-implicit-bias-review/> [accessed 19 February 2018]

Stanovich, Keith. 2009. 'Distinguishing the Reflective, Algorithmic, and Autonomous Minds: Is It Time for a Tri-Process Theory?', *In Two Minds: Dual Processes and Beyond*, ed. by Jonathan Evans and Keith Frankish (Oxford: Oxford University Press), pp. 55–88

Stanovich, Keith, and Richard West. 2003. *Evolutionary Versus Instrumental Goals: How Evolutionary Psychology Misconceives Human Rationality*, ed. by David E. Over (New York: Psychology Free Press (Tayor and Francis Group) Also available <https://semioticon.com/virtuals/imitation/kstanovich_paper.pdf> [accessed 9

December 2020)

Steward, Helen. 2011. 'Moral Responsibility and the Concept of Agency', *Free Will and Modern Science* (Oxford: Oxford University Press), pp. 141–57

———. 2012a. *A Metaphysics for Freedom* (Oxford University Press) <https://doi.org/10.1093/acprof:oso/9780199552054.001.0001>

———. 2012b. 'The Metaphysical Presuppositions of Moral Responsibility', *The Journal of Ethics*, 16.2: 241–71 <https://doi.org/10.1007/s10892-012-9127-5>

Steward, Helen, and Michael Hauskeller. 2020. *Agency* <https://www.academia.edu/42056230/Agency> or via Michael Hauskeller's area under the heading 'Conversations' <https://liverpool.academia.edu/MichaelHauskeller> [accessed 17 July 2020]

Strack, Fritz, and Roland Deutsch. 2004. 'Reflective and Impulsive Determinants of Social Behavior', *Personality and Social Psychology Review*, 8: 220–47

Strawson, Galen. 1994. 'The Impossibility of Moral Responsibility', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 75: 5–24

Strawson, Peter. 1962. 'Freedom and Resentment', *Proceedings of the British Academy,* 48: 187–211

———. 2008. 'Freedom and Resentment', *Freedom and Resentment and Other Essays*, First edn. (Oxford: Routledge), pp. 1–28

Stump, Eleonore. 1999. 'Alternative Possibilities and Moral Responsibility: The Flicker of Freedom', *The Journal of Ethics*, 3.4: 299–324 <https://www.jstor.org/stable/25115622>

Stump, Eleonore, and Norman Kretzmann. 1981. 'Eternity', *The Journal of Philosophy*, 78.8: 429–58 <https://doi.org/10.2307/2026047>

Sullivan-Bissett, Ema. 2019. 'Biased by Our Imaginings', *Mind and Language*, 34.5: 627–47 <https://doi.org/10.1111/mila.12225>

Swinburne, Richard. 1993. *The Coherence of Theism*, Revised edn. (Oxford: Clarendon)

Tauber, Alfred I. 2010. *Freud the Reluctant Philosopher*, Kindle edn. (Oxford: Princeton University Press)

Teige-Mocigemba, Sarah, Karl Christoph Klauer, and Jeffrey W. Shermen. 2010. 'A Practical Guide to the IAT and Related Tasks', *Handbook of Implicit Social Cognition*, ed. by Bertram Gawronski and B. Keith Payne (New York: Guilford Press), pp.

117–39

Tesser, Abraham. 1978. 'Self-Generated Attitude Change', *Advances in Experimental Social Psychology*, 11: 289–338 <https://doi.org/10.1016/S0065-2601(08)60010-6>

Thomas, Nigel. 2019. 'Mental Imagery', *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)* (University of Stanford), <https://plato.stanford.edu/entries/mental-imagery/> [accessed 24 August 2020]

Timpe, Kevin. 2003. 'Free Will', *Internet Encyclopedia of Philosophy* <https://doi.org/10.4324/9780415249126-V014-2> [accessed 20 November 2020]

Tversky, Amos, and Daniel Kahneman. 1974. 'Judgment under Uncertainty: Heuristics and Biases (1974)', *Science, (New Series)*, 185.4157: 1124–31 <http://www.jstor.org/stable/1738360>

Vargas, Manuel R. 2008. *Revisionism about Free Will: A Statement and Defense*, (The University of San Francisco USF Scholarship: A Digital Repository @ Gleeson Library) <https://doi.org/10.1007/sll098-009-9366-x>

———. 2017. 'Implicit Bias, Responsibility, and Moral Ecology', *Oxford Studies in Agency and Responsibility Volume 4,* First edn. (Oxford: Oxford University Press), pp. 219–47

Vihvelin, Kadri. 2008. 'Foreknowledge, Frankfurt, and the Ability to Do Otherwise: A Reply to Fischer', *Canadian Journal of Philosophy*, 38.3: 343–72

———. 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*, (Oxford: Oxford University Press (Oxford on Line)) <https://doi.org/10.1093/acprof:oso/9780199795185.001.0001>

Walker, Margaret. 1993. 'The Virtues of Impure Agency', *Moral Luck*, ed. by Daniel Statman (New York: State University of New York), pp. 235–50

Walther, Eva. 2002. 'Guilty by Mere Association: Evaluative Conditioning and the Spreading Attitude Effect', *Journal of Personality and Social Psychology*, 82.6: 919–934

Washington, Natalia, and Daniel Kelly. 2016. 'Who's Responsible for This?', *Implicit Bias and Philosophy Vol 2*, ed. by Michael Brownstein and Jennifer Saul (Oxford: Oxford University Press), pp. 11–36

Watson, Gary. 2013. 'Responsibility and the Limits of Evil: Variations on a

Strawsonian Theme', *The Philosophy of Free Will*, ed. by Paul Russell and Oisin Deery (New York: Oxford University Press), pp. 84–113

Weil, Simone. 1965. 'The Iliad, or the Poem of Force', *Chicago Review*, 18.25 <http://biblio3.url.edu.gt/SinParedes/08/Weil-Poem-LM.pdf> [accessed 20 March 2017]

Widerker, David, and Stewart Goetz. 2013. 'Fischer against the Dilemma Defence: The Defence Prevails', *Analysis*, 73.2: 283–95 <https://doi.org/10.1093/analys/ant013>

Wieland, Jan Willem. 2017. 'Introduction', *Responsibility: The Epistemic Condition*, ed. by Jan Willem Wieland and Philip Robichaud (Oxford: Oxford Scholarship Online), pp. 1–36 <https://doi.org/10.1093/oso/9780198779667.001.0001>

Wierenga, Edward. 2011. *Omniscience* (Oxford: Oxford University Press) <https://doi.org/10.1093/oxfordhb/9780199596539.013.0007>

Williams, Bernard. 1981. *Moral Luck* (Cambridge: Cambridge University Press)

———. 2008. *Shame and Necessity*, (USA: University of California Press), pp. 254

Wilson, Timothy, Samuel Lindsey, and Tonya Schooler. 2000. 'A Model of Dual Attitudes', *Psychological Review*, 107: 101–26 <https://www.researchgate.net/profile/Timothy_Wilson21/publication/12628954_A_Model_of_Dual_Attitudes/links/576ee8eb08ae0b3a3b79ce6d.pdf>

Wolf, Susan. 1993. *Freedom Within Resean* (New York: Oxford University Press)

———. 2001. 'The Moral of Moral Luck', *Philosophic Exchange*, 31.1: 1–16

———. 2003. 'Sanity and the Metaphysics of Responsibility', *Free Will (Oxford Readings in Philosophy)*, Second edn., ed. by Gary Watson (Oxford: Oxford University Press), pp. 372–87

Zagzebski, Linda. 1985. 'Divine Foreknowledge and Human Free Will', *Religious Studies*, 21.21: 279–98 <https://doi.org/10.1017/S0034412500017406>

———. 2017. 'Foreknowledge and Free Will', *Stanford Encyclopaedia of Philosophy* (Stanford University) <https://plato.stanford.edu/entries/free-will-foreknowledge/> [accessed 27 June 2017]

Zimmerman, David. 2002. 'Reasons-Responsiveness and Ownership of Agency: Fischer and Ravizza's Historicist Theory of Responsibility', *The Journal of Ethics*, 6.3: 199–234

Appendix A

Mental Representations and Biased Behaviour

In this Appendix I consider the nature of mental representations responsible for biased behaviour and describe a position that is not built upon propositional attitudes or mere associations, rather, drawing on both positions, it is based on the concept of mental imagery. Bence Nanay points out, '… what these biasing representations are is crucial not just out of theoretical interest. If we want to try to eliminate implicit bias, very different procedures would be needed depending on what these biasing representations are' (2020). Continuing with Nanay, straightforward examples of associationist and propositional approaches to bias representation are given below. First, association:

> You have probably seen more female caregivers than male caregivers. And, following the mechanism of classic conditioning, you formed an association between being a caregiver and being a woman. One way to think about associations is as some kind of connection strength in your mind between the concept of being a caregiver and the concept of being a woman. When one concept is activated, the other one is highly likely to be also activated. So, when you hear someone talk about a caregiver, this gives rise to you thinking of a woman. Association is supposed to be quick, not under our voluntary control and, according to many … symmetrical (it goes both from caregiver to woman and vice versa). (2020)

Alternatively, if the underlying biasing representation is a propositional attitude (typically a belief):

> So, you have a propositional attitude that caregivers are (likely to be) women. And it is this propositional attitude that explains your biased behavior. In the case of propositional attitudes, the relation between being a caregiver and being a woman is not symmetrical. The propositional attitude that caregivers are (or tend to be) women is different from the propositional attitude that women are (or tend to be) caregivers. (2020)

There are empirical examples, (Mandelbaum 2016; Nanay 2020), that claim to illustrate propositional *and* associative approaches to be problematic. For example, if biasing representations are associations, then conditioning could counteract them, but it is claimed that conditioning does not counteract bias, hence biasing representations are not associations. If biasing representations are propositional in nature then implicit bias should be sensitive to logical form, however, it is claimed, for example, that the sentence 'it is not true that old people are bad drivers' may actually strengthen implicit bias by the association of 'old people' and 'bad drivers' rather than reducing bias.[220] As Nanay notes, 'if the biasing representation were a propositional attitude (presumably a propositional attitude about old people driving badly), then exposure to a sentence that denies the content of this propositional attitude should not strengthen the implicit bias' (2020). Similar examples are given that lead to a general conclusion that a way must be found that essentially takes forward what is true from both accounts into a new paradigm of bias representation.

Nanay claims that such a way forward is possible, based on mental imagery.[221] Mental imagery appears to the subject as perceptual experience, however, unlike actual perceptual experience there is an absence of external stimuli. 'Visual mental imagery, the most discussed variety, was thought to be caused by the presence of picture-like representations (mental images) in the mind, soul, or brain, but this is no longer universally accepted' (Thomas 2019). Such images may be in some sense a construction based on experience, or of some imagined future experience that may be desired or feared. How does this concept help? The idea is developed by Nanay as follows;

- It is claimed that experimental results about implicit bias show that the biasing representation, what exists or mediates between a stimulus and resulting behaviour, is sensitive to semantic content and insensitive to logical form (controversial). [222]

[220] See Deutsch and Strack (2010: 64) for discussion of association – proposition accounts within theories of implicit social cognition. (Deutsch and Strack's *integrated* model is the prime model used during critique of semicompatibilism in Part III). There are many sources for discussion of association – proposition accounts; Bertram Gawronski and Laura A. Creighton, *Dual Process Theories* (2013) is excellent.

[221] See also Ema Sullivan-Bissett, *Biased by Our Imaginings* (2019).

[222] I have not reproduced examples that lead to this claim, such description may be found in Section III, *Implicit Bias as Mental Imager* (Nanay 2020).

- Neither associations nor propositions satisfy both requirements. The classic accounts of implicit bias are problematic; associations are not sensitive to semantic content and propositional attitudes are not insensitive to logical form.
- Mental imagery satisfies both requirements, sensitive to semantic content unlike associations but insensitive to logical form unlike propositions.

I will say more about mental imagery, describing the claim that this concept satisfies the requirement of representation as sensitive to semantic content and insensitive to logical form. Nanay does not go into details of the debate about the format of perception but moves forward noting the assumption that representation has imagistic (precise images) non-propositional content. Sullivan-Bissett (2019) has much more to say, describing one (of two possible) non associative processes. Sullivan-Bissett does not have the same project as Nanay, (the unconscious imagination perspective of Sullivan-Bissett's Paper is different from mental imagery), but the following sheds light on the role of mental imagery in the context of bias:

> First, when presented with a woman-stimulus, one could imagistically unconsciously imagine a weak woman. In this case, the implicit bias is identical to a single instance of unconscious imagistic imagining, rather than being the association between two unconscious mental images. (2019: 637)

Following closely Nanay's text; as noted, mental imagery is not a propositional attitude (it has imagistic content), therefore mental imagery does not enter into inferences. However, not all *content-sensitive* transitions between mental states are inferences and there can be content-sensitive transitions between mental states with imagistic content. An example is given of mental imagery leading to other mental processes in a way that is content-sensitive. For example, in a gift-wrapping task, the size of the wrapping paper to be cut from the roll is estimated by looking at the gift and the paper in a content-sensitive manner by exercising mental imagery (2020). The key point is, the perspective of mental imagery describes content-sensitive transition between mental imagery and other mental processes not mediated by beliefs, other propositional attitudes or association. There are further considerations that motivate the role of mental imagery within implicit bias, particularly the idea that mental imagery may be conscious or unconscious. Ema Sullivan-Bissett, *Biased by Our Imaginings* (2019), argues specifically that implicit biases are constituted by unconscious imaginings. Nanay summarises the properties of mental imagery within the context of implicit bias as affectively charged and action-oriented involuntary mental imagery that may be conscious or unconscious, subject to empirical and philosophical resolution, both states being allowed within the model.

There is a reasonable suspicion that the mental imagery approach has not resolved the claimed problems with the associationist and propositionalist accounts. There is a transition between trigger and mental image, and mental image and action. What is the nature of this transition? Such transitions may be explained either in associationist or propositionalist terms. Nanay accepts this criticism but claims progress has been made by arguing the current question concerning how a perceptual state gives rise to mental imagery and how mental imagery turns into behavior is quite a different question, and is a progressive step, from the question what kind of connection is there between concepts that lead to biased behavior? I take Nanay's point but question the actual size of the step. That said, there is significant success claimed in counteracting implicit bias using various programs of manipulation of the subjects' mental imagery. The well-known process of trying to visualise ourselves as members of another racial group or gender is claimed to reduce the consequences of implicit bias when in the company of these groups; the imagery-involving procedure is claimed to be among the most efficient ways of reducing implicit bias (Nanay 2020). Assuming such bias mitigating strategies are successful, it is difficult for the propositionalist to explain how this is possible. It is difficult to explain how a strategy involving a perceptual process could have any effect on a biasing representation that is a propositional attitude. Similarly, it is claimed that imagery-involving mitigation procedures are more successful than techniques that typically assume associationism, (repeated exposure to perceptual stimuli that goes against the association). The associationist would have difficulty explaining this. A good case is made for the mental imagery perspective; given that Nanay admits a slight weakness in his position on the nature of transitions I assume this will be the subject of future work.

Alternatives to the classic association – proposition positions on biasing representation have been described, based closely on Bence Nanay *Implicit Bias as Mental Imager* (2020), also Sullivan-Bissett *Biased by Our Imaginings* (2019). One aim of this Thesis is to present a broad description of implicit bias and I believe providing an outline of another alternative to the fundamental association – proposition positions on biasing representation contributes to that objective.

Appendix B

Agent Causation

The terms 'agent' and 'agency' have been used throughout this Thesis. It is important to comment on agency specifically, as the concept is deeply linked with discussion of free will and responsibility. I will mention the standard conception of agency and standard theory of action, then briefly outline three metaphysical frameworks of agency and close with *A Metaphysics for Freedom* (Steward 2012a) and a brief outline of compatibilism, agency and emergentism. The breadth and depth of these subjects is clearly beyond the scope of an Appendix, nonetheless, an outline is attempted to give a sense of the importance of agency and connecting themes within discussion of free will and responsibility.

The Stanford Encyclopedia of Philosophy (Schlosser 2019), drawing on G.E.M Anscombe *Intention* (1957) and D. Davidson *Actions, Reasons, and Causes* first published 1963 and reprinted in *Essays on Actions and Events* (Davidson 1980: 3-20), describes the standard conception of agency and the standard theory of agency as follows:

> … a being has the capacity to exercise agency just in case it has the capacity to act intentionally, and the exercise of agency consists in the performance of intentional actions and, in many cases, in the performance of unintentional actions (that derive from the performance of intentional actions). Call this the standard conception of agency. The standard theory of action provides us with a theory of agency, according to which a being has the capacity to act intentionally just in case it has the right functional organization: just in case the instantiation of certain mental states and events (such as desires, beliefs, and intentions) would *cause* the right events (such as certain movements) in the right way. According to this standard theory of agency, the exercise of agency consists in the instantiation of the *right causal relations between agent-involving states and events*. (added emphasis Schlosser 2019)

On this view, agency is the exercise of the capacity to perform intentional actions and the process of exercising the capacity to perform intentional actions, (i.e., being an agent), is an instance of the right *causal* relations between agent-involving states and events. The standard conception and standard theory have been subject to substantial and varied scrutiny which cannot be adequately presented here. However, one general area of criticism concerns the presence of mental states within this model. Criticism is usually based around one or more of the following three claims. First, controversially, it is claimed there are non-human beings that show agency and do not possess representational mental states; therefore, a mental states-based explanation is inadequate. Second, there are many instances of human agency that can and should be explained without ascription of representational mental states such as actions that do not involve deliberation. Therefore, a purely mental states-based explanation of agency is again inadequate. Third, more radically, *all* instances of agency it is claimed can and should be explained without ascription of representational mental states. From earlier discussion of the control of implicit bias, one is reminded of the substantial ideas of embodied and embedded cognition[223] in the sense that description of human cognition in terms of mental states alone is far from complete.

Having outlined the standard model of agency and mentioned one general critical approach concerning mental states, I will outline three metaphysical frameworks that try to explain the nature of agency and the relation between agents and actions drawing on section 3.1 of Markus Schlosser's Stanford Encyclopedia of Philosophy article *Agency* (2019). First, the event-causal framework: The agent's role in exercising agency is characterised in terms of causation by the agent's mental states. In a world where determinism is true, such mental states would be the result of previous events stretching

[223] Embodied cognitive science appeals to the idea that cognition deeply depends on aspects of the agent's *body* other than the brain. Without the involvement of the body in both sensing and acting, thoughts would be empty, and mental affairs would not exhibit the characteristics and properties they do. Work on embedded cognition, by contrast, draws on the view that cognition deeply depends on the natural and social environment. The thesis of extended cognition is the claim that cognitive systems themselves extend beyond the boundary of the individual organism. Features of an agent's physical, social, and cultural environment can do more than distribute cognitive processing: they may well partially constitute that agent's cognitive system (Schlosser 2019). See also *An Introduction to Implicit Bias* (Beeghly, Erin, and Madva (eds.) 2020) and *Embodied Freedom* (A Draft Paper by Jeffrey Pannekoek, University of Tennessee, <https://utk.academia.edu/JeffreyPannekoek>) for detailed discussion of these ideas.

back into the past, and together with fixed laws of nature would result in actions that conventionally would not be considered free actions.

Second, the agent-causal framework: Agency is characterised in terms of a kind of emergent substance-causation, where causation by the agent is understood as a persisting substance; a kind of capability possessed by an agent. (To clarify, emergentism is a form of ontological materialism (or physicalism), the view that the contents of the world are exhausted by matter (Kim 2009: 10). Schlosser's description of agent causation as 'a kind of *emergent* substance-causation' I believe is not committed to any view on the nature of the 'emergent substance'). The agent's role in the exercise of agency may be understood in terms of the exercise of an *irreducible* novel emergent agent-causal power. On this view, there is sufficient elbow room within a determined world for free actions to be initiated. There *will* be a causal chain, but the origin of that chain is the causal agent. Such a capability is problematic; the possibility of a material entity, i.e., a person, having such an emergent, (yet essentially material), property of downward causation acting on, yet independent of, the rest of the world is difficult to understand. Agent causation is an incompatibilist (personal freedom and a determined world are incompatible), libertarian position, (freedom *is* present and determinism inactive *at the origin* of the causal process that leads to actions by the agent).

Third, the volitionist approach: Agency is explained in terms of acts of the will, usually called 'volitions'. On this view, volitions are the source of agency. Volitions themselves are entirely uncaused and they are sui generis acts: they are acts by virtue of their intrinsic properties, not because of some extrinsic or relational property (such as having the right causal history).

At the time of writing (July 2020), the most accepted position is the event-causal framework (Schlosser 2019). One plausible reason for greater acceptance is its relative simplicity, not requiring a mysterious emerging 'something' that is active independent of event-causation. However, the familiar problem of *control* over actions when all intentional actions are explained in terms of event-causation is clearly present. The volition model also has control issues in the sense that agent control of issuing behaviour appears absent as the agent is just the subject of volitions having no input into what is willed. The agent causal model previously described, by contrast, has influencing factors within the causal chain that originate within the agent. Factors that may include *reasons*, plans and the agent's disposition. (See below mention of Timothy O'Connor's article *Agent-Causal Power* (2013)). Of course, it could be argued that such influences may be freedom diminishing, being the result of factors over which the agent has no control such as constitutional luck. Further, if volitions ultimately have brain states as their origin, then acting independently of causation appears impossible. It is not possible here to make a comparative analysis of these three positions, however, it is valuable to continue by briefly considering one philosopher's analysis of agency from a metaphysical perspective, one that discusses agency considering responsibility within a deterministic world. I will now mention Helen Steward *Moral Responsibility and the Concept of Agency* (2011) before concluding with comments on compatibilism and agency, and further clarification of emergence.

In a conversation between Helen Steward and Michael Hauskeller (Steward and Hauskeller 2020), Hauskeller, from an agent-causal perspective, summarises the problem of agency very well:

> It seems to require something that is in fact conceptually far more difficult to grasp than the suspension of the principle of causality. What we need to get our heads around is not causation or its absence, but the possibility of self-causation. A true agent, it seems, is, like the God of some philosophers, causa sui, their own cause. Is this something we can really make sense of? In order to understand agency, we need to carve out a space for the agent to genuinely settle things: things that are neither uncaused nor *causally* determined by anything but the agent herself. (added emphasis 2020)

Later, Steward comments:

> Surely no event in nature can just begin! Of course, actions require underlying neurological activity, and to that extent I would agree that we need some sort of account of how the causal principles which govern the underlying neurological hardware are to be rendered consistent with the idea that whether or not an action occurs is up to its agent at the time of action. Ultimately, it seems to me, what we need to understand is not self-causation, exactly, but rather *whole-part* causation. (added emphasis Steward and Hauskeller 2020)

Steward's position is essentially incompatibilist, but from an unusual perspective within most determinism – free will – responsibility literature. It is claimed that the reason responsibility and determinism are incompatible is not moral in the sense that determinism denies an agent the responsibility delivering power of doing otherwise, rather, determinism as it is usually understood, is inconsistent with *agency*, which is a necessary condition of moral responsibility, i.e., agency incompatibilism. A libertarian position is developed by Steward whereby 'higher animals do have powers to make certain things happen, particularly changes to the distribution and arrangement of their own bodily parts in ways not merely dictated by the past and the laws' (2011: 8). It is only possible to mention here that Steward develops a detailed concept of agency incompatibilism, within a detailed metaphysical framework.[224] A model of top-down causation is established, suggesting a natural, (without need of an additional substance), and distinctive emergent *agent causal* power over lower-level occurrences within a biological hierarchy that is the animal itself (following Steward 2012a: 24).

Having given a gloss of the agency debate, I turn specifically to compatibilism, the consequence argument, further mention of agent-causation and implicit bias. Compatibilism in terms of an agent as a source or origin of free (and responsible) actions can be described as follows: The notion of a part of ourselves acting in an important sense independent from 'external' determining factors is at once both easy and difficult to understand. It is easy because, as previously mentioned, the experience of making decisions feels completely free; deciding whether to continue typing or to stop work and make a drink *feels* to the one that chooses a completely free exercise of *their* agency. It is difficult, because looking at the world provides no such intimate first person sense of freedom, and the possibility of determined actions and behaviour beyond ourselves is not easily dismissed (following Thomas Nagel (2003: 229).

The *consequence* argument discussed previously, (see page 37), is the first of two basic arguments used by incompatibilists to support their position, outlined briefly again below:

1. There is nothing we can now do to change the past.
2. There is nothing we can now do to change the laws of nature.

[224] See all related works by Helen Steward, particularly *Moral Responsibility and the Concept of Agency* (Steward 2011), *The Metaphysical Presuppositions of Moral Responsibility* (2012b) and *A Metaphysics for Freedom* (Steward 2012a).

3. There is nothing we can now do to change the past and the laws of nature.
4. *If* determinism is true, our present actions are the necessary *consequences* of the past and the laws of nature. (That is, it must be the case that, given the past and the laws of nature, our present actions occur).
5. Therefore, there is nothing we can now do to change the fact that our present actions occur.

In other words,[225] determinism and free choice are incompatible. This is of cause controversial, generating considerable debate concerning the nature of time, causation, and so on. The second basic incompatibilist argument, to be considered here, concerns the agent as source or origin of free (and responsible) actions and is called the *origination* argument. Intuitively, the notion of free action sustained by its origin within the agent seems, based on our experience of the world, very plausible, but the origination argument highlights some difficulties. Typically, the argument is as follows (Timpe 2003):

1. An agent acts with free will only if she is the originator (or ultimate source) of her actions.
2. If determinism is true, then everything any agent does is ultimately caused by events and circumstances outside her control.
3. If everything an agent does is ultimately caused by events and circumstances beyond her control, then the agent is not the originator (or ultimate source) of her actions.
4. Therefore, if determinism is true, then no agent is the originator (or ultimate source) of her actions.
5. Therefore, if determinism is true, no agent has free will.

There are many ways to counter this incompatibilist argument; the strategy described here looks at the first premise and develops a perspective where agents *do* act with free will because they *can be* the originator (or ultimate source) of their actions.[226]

[225] Literature concerning these concepts is vast, but previously mentioned works by Helen Steward provide clear and detailed exposition of the issues.

[226] An alternative strategy for compatibilists is rejection of premise 1, arguing that free action *is* possible even if the agent is *not* the originator (or ultimate source) of her actions. For example, with Frankfurt, if

Title: Does Implicit Bias Threaten the Semicompatibilist Position on Free Will and Responsibility?
Appendix B
Agent Causation

Timothy O’Connor’s article *Agent-Causal Power* (2013) is important within the agency debate. The fundamental claim is that ontologically emergent powers confer ontologically primitive causal power upon an agent, so provide the necessary form of control for freedom of action. Such causal power grants the agent a particular power to cause a ‘certain type of event within the agent: the coming to be of a state of intention to carry out some act, thereby resolving a state of uncertainty … ’ (O’Connor 2013: 233). An agent-causal event occurs in the *presence* of ‘motivational states’ that are present prior to *the agent* causing an intention to act; O’Connor notes that the influence of ‘motivational states’ on the nature of the intention may be appreciable (2013: 235). Following this description, O’Connor makes a crucial point; such motivational states could include sufficient reasons for acting ‘of which I am entirely *unconscious*’ (2013: 235). Clearly this suggests the possibility of implicit bias as a contributor to the collection of motivators, but not in a direct causal role. If implicit bias is an integral part of a process with freedom at its heart, then the presence of implicit bias does not appear to fatally threaten freedom (and responsibility) within an agent-causal account of compatibilism. However, before making claims of this sort there is more work to do. While O’Connor notes that motivational states could include sufficient *reasons* for acting ‘of which I am entirely unconscious’, he is not saying that the *whole* process prior to action is unconscious. There is awareness of inclination, but the reasons behind that inclination may not be fully known or may be completely unknown and for this reason freedom is diminished because reasons are not necessarily available to rational examination. When a particular action is influenced by unconscious factors it is nonetheless a free action. O’Connor’s claim is clear on this point; acting from unconscious reasons diminishes freedom, but nonetheless ‘it remains open to me to undertake the action or not, I exhibit the … self-determination that is the core element of freedom of the will’ (2013: 236). In the case of implicit bias, an agent is aware of their *inclination*, i.e., to avoid contact with someone of a particular ethnicity, but are not consciously aware of the influencing bias that lies behind the conscious inclination. Awareness of an inclination to act in a particular way is crucial, facilitating the possibility to act differently with control and

1st order volitions and 2nd order desires mesh appropriately then free action is possible without the agent as originator or ultimate source.

responsibility; awareness and control of issuing behaviour without awareness or control concerning the possession of implicit biases.

Before summarising, some final comments concerning emergence. Clearly, if the claim that agents act freely is based upon an emergent property, the actual phenomenon of emergence calls for further clarification.

The key point is the notion of reality as layers in a hierarchical structure, where the bottom level is occupied by the most basic arrangements of matter and at increasingly higher levels complexity increases but *no* additional substance comes into play. Hence, emergence or emergentism is a form of ontological materialism (or physicalism); the world is comprised of matter, nothing more, just matter in varying degrees of complexity. As complexity increases, at higher levels of the hierarchy properties of living organisms emerge, with humans (and probably some animals) having qualities such as abstract thought and rationality emerging at the highest level of complexity, (given current development). Crucially, some of the properties at the higher levels are *unique* to that level and are not currently[227]explainable or predictable based on knowledge of lower, less complex arrangements of matter. Expressed more succinctly, 'some of the properties characterizing entities of that (higher) level are *irreducible* to the properties at the lower levels' (Kim 2009: 12). Kim clarifies emergentism and reductionism:

> … emergent properties are those that are not reducible to lower-level properties, whereas non-emergent, or resultant, properties are reducible. Emergentism and reductionism, then, are the two options regarding the metaphysical question concerning the inter-level relationship of properties. […] Emergentism with respect to a given level is the claim that some of the properties characterizing entities of that level are irreducible to the properties at the lower levels. Reductionism denies this, claiming that all properties at that level are reducible. (Kim 2009: 12)

It is clear that an important question concerns how properties that characterise one level interact with those of other levels, particularly with those of lower levels? This question suggests the mind-body problem i.e., the nature or even possibility of interaction

[227] See *Studies in the Logic of Explanation* for careful exposition of this point, i.e., that 'no explanation, in terms of microstructure theories, is available *at the present* for large classes of phenomena studied in biology and psychology' (Hempel and Oppenheim 1948).

between mind and body, (however, unlike emergent properties, mind is not a material or physical substance). More generally, how is it possible for the emergent mind having the novel property of consciousness to exercise or assert its causal powers in a world largely occupying a lower position within an emergentist hierarchy (following Kim 2007: 7)? Without mental causation there is no human action or agency, without consciousness our whole conception of ourselves as thinking, reflective human beings, is lost.

Emergentists often claim that emergent properties can exert their own distinctive causal powers within the material world; consciousness, thought and rationality having causal power is something most people would endorse based on experience without hesitation, however, the claim of 'downward' causation within the emergentist hierarchy remains controversial.[228]

Summarising, I have outlined the standard conception of agency and standard theory of action, then briefly outlined three metaphysical frameworks of agency and looked at compatibilism, agency and emergentism. I find the idea of agent causation very plausible as an emergent human (and some, perhaps many, nonhuman animal) property within an incompatibilist/libertarian model of free will. While the irreducible nature of emergent human agency is currently (and may always remain) mysterious, the attractiveness of ontological materialism is undeniable. Human consciousness, while similarly mysterious when described in terms of such an emergent model, is nonetheless undeniable.

[228] Detailed criticism of the emergentist position may be found in Carl G. Hempel and Paul Oppenheim, *Studies in the Logic of Explanation* (1948: 148-152) and the classic C. D. Broad, *The Mind and its Place in Nature* (1925: 43-94). Kadre Vihvelin, *Causes, Laws, and Free Will: Why Determinism Doesn't Matter* (2013) gives a particularly clear and detailed account of agent causation. See also all related Papers by Helen Steward for detailed account of agent causation.

Appendix C

Plot Summary

Sophocles: Oedipus Rex[229] c429 BCE

- King Oedipus of Thebes sends his brother-in-law Creon to find the cause of the mysterious plague that has struck the city. Creon reports that the plague will be lifted if the man who killed the former king, Laius, is brought to justice.

- Queen Jocasta does not believe Tiresias when he says Oedipus is the murderer. Once, an oracle told her that their child would kill her husband, and because she believes that has not come true, she does not believe Tiresias.

- To prevent her child from killing her husband, Jocasta left the baby to die on the side of the road. Oedipus suspects that he was that abandoned baby. When he first came to Thebes he met and killed a man on the road who turned out to be Laius, his father. He then met and married the widowed Jocasta, his own mother.

- A messenger and a servant confirm the tale. Jocasta hangs herself out of shame. Oedipus discovers her body and uses the pins of her brooches to stab out his own eyes.

Summary

When Thebes is struck by a plague, the people ask King Oedipus to deliver them from its horrors. Creon, the brother of Jocasta, Oedipus's queen, returns from the oracle of Apollo and discloses that the plague is punishment for the murder of King Laius, Oedipus's immediate predecessor, to whom Jocasta was married. Creon further discloses

[229] Source: 'Oedipus Rex – Summary', *Critical Survey of Literature for Students*, Ed. Laurence W. Mazzeno, eNotes.com Inc. 2010 <http://www.enotes.com/topics/oedipus-rex#summary-summary-summary-the-story> [accessed 26 May 2017].

that the citizens of Thebes need to discover and punish the murderer before the plague can be lifted. The people mourn their dead, and Oedipus advises them, in their own interest, to search out and apprehend the murderer of Laius.

Asked to help find the murderer, Tiresias the ancient blind seer of Thebes tells Oedipus that it would be better for all if he does not tell what he knows. He says that coming events will reveal themselves. Oedipus rages at the seer's reluctance to tell the secret until he goads the old man to reveal that Oedipus is the one responsible for Thebes's afflictions because *he* is the murderer, and that he is living in intimacy with his nearest kin. Oedipus accuses the old man of being in league with Creon, whom he suspects of plotting against his throne, but Tiresias answers that Oedipus will be ashamed and horrified when he learns the truth about his true parentage. Oedipus defies the seer, saying he will welcome the truth if it frees his kingdom from the plague. Oedipus threatens Creon with death, but Jocasta and the people advise him against doing violence on the strength of rumour or momentary passion. Oedipus yields, but he banishes Creon.

Jocasta, grieved by the enmity between her brother and Oedipus, tells her husband that an oracle informed King Laius that he would be killed by his own child, the offspring of Laius and Jocasta. Jocasta assures Oedipus that this could not happen because the child was abandoned on a deserted mountainside soon after birth. When Oedipus hears further that Laius was killed by robbers at the meeting place of three roads and that the three roads met in Phocis, he is deeply disturbed and begins to suspect that he is, after all, the murderer. He hesitates to reveal his suspicion, but he becomes increasingly convinced of his own guilt.

Oedipus tells Jocasta that he believed himself to be the son of Polybus of Corinth and Merope until a drunken man on one occasion announced that the young Oedipus was not really Polybus's son. Disturbed, Oedipus consulted the oracle of Apollo, who told him he would sire children by his own mother and that he would kill his own father. After he left Corinth, at a meeting place of three roads, a man in a chariot offended Oedipus. He killed the man and all his servants but one. From there he went on to Thebes, where he became the new king by answering the riddle of the Sphinx. The riddle asked what went on all fours before noon, on two legs at noon, and on three legs after noon. Oedipus answered, correctly, that human beings walk on all fours as an infant, on two legs in their prime, and with the aid of a stick in their old age. With the kingship, he also won the hand of Jocasta, King Laius's queen.

Oedipus summons the servant who reported King Laius's death, but he awaits his arrival fearfully. Jocasta assures her husband that the entire matter is of no great consequence, that surely the prophecies of the oracles will not come true.

A messenger from Corinth announces that King Polybus is dead, and that Oedipus is his successor. Polybus died of natural causes, so Oedipus and Jocasta are relieved for the time being. Oedipus tells the messenger he will not go to Corinth for fear of siring children by his mother, Merope.

The messenger goes on to reveal that Oedipus is not the son of Polybus and Merope but a foundling whom the messenger, at that time a shepherd, took to Polybus. The messenger relates how he received the baby from another shepherd, who was a servant of the house of King Laius. At that point Jocasta realizes the dreadful truth. She does not wish to see the old servant who was summoned, but Oedipus desires clarity regardless of the cost. He again calls for the servant. When the servant appears, the messenger recognizes him as the herder from whom he received the child years earlier. The old servant confesses that King Laius ordered him to destroy the boy but that out of pity he gave the infant to the Corinthian to be raised as his foster son.

Oedipus, now all but mad from the realization of what he did, enters the palace and discovers that Jocasta hanged herself by her hair. He removes her golden brooches and with them puts out his eyes so that he will not be able to see the results of the horrible prophecy. Then, blind, bloody and miserable, he displays himself to the Thebans and announces himself as the murderer of their king and the defiler of his own mother's bed. He curses the herder who saved him from death years before.

Creon, returning, orders the attendants to lead Oedipus back into the palace. Oedipus asks Creon to have him conducted out of Thebes where no man will ever see him again. He also asks Creon to give Jocasta a proper burial and to see that the children of the unnatural marriage should be cared for and not be allowed to live poor and unmarried because of the shame attached to their parentage. Creon leads the wretched Oedipus away to his exile of blindness and torment.

~~~~~~
~~~~~~