Autonomy in Political Liberalism

Luke Armstrong
MA, MRes

Submitted in fulfilment of the requirements of the Degree of Doctorate of Philosophy

School of Social and Political Sciences, College of Social Sciences
University of Glasgow

October 2020

Abstract

John Rawls intends his doctrine of political liberalism to be free of metaphysical commitments. For this reason, Rawls' conception of autonomy is not supposed to be dependent on the truth of any one metaphysical theory of human nature. Instead, Rawls states that people are rationally autonomous – able to develop their moral powers – and fully autonomous – able to act from a sense of justice – positing that we can accept these conceptions of autonomy whatever our other philosophical commitments.

Though autonomy is not Rawls' central concern, I argue that, nevertheless, full autonomy is integral to Rawls' theory of justice. Without fully autonomous citizens, there is no one to realise justice as fairness. Through his way of thinking about human nature, Rawls assumes that people will be motivated to become fully autonomous. Through the possession of the moral powers, and the witnessing of the just workings of society's institutions, a person will wish to act as a fully autonomous citizen should. I argue against making this assumption. Instead, I evaluate Rawls' conception of the person against data in neuroscience and psychology, and thereby articulate two central concerns. First, a person's moral psychology is largely dependent on her upbringing. If a person's upbringing has not instilled in her a sense of the importance of fairness, there is little hope of her becoming fully autonomous in the way Rawls imagines. Second, in the neuroscience of free will, evidence suggests that we are much less in control of our thought processes than is traditionally imagined. If we are not in full control of our thought processes, there may be little we can do to prevent our being influenced by ideas that undermine justice as fairness.

This is why I argue that, to realise political liberalism in the way Rawls formulates it, a system of moral education is necessary; a system that is much more demanding than that imagined by Rawls. It cannot be assumed that people are inherently predisposed to the value of justice as fairness, nor can it be imagined that the majority of people would reject alternative doctrines through their capacity to regulate their thoughts. A strong prior commitment to justice as fairness is therefore necessary. If political liberalism is to be realised in the way Rawls imagines, people must be educated in the importance of justice as fairness, with the aim of such a system of education being the development of fully autonomous citizens.

Table of Contents

Acknowledgements

Many thanks are owed to my supervisors, Carl Knight and Ben Colburn, who have been supportive of this project throughout. They read each draft thoroughly, offering me incisive comments, and encouraged me to strengthen my arguments. Their responses to my work often helped me to clarify my own thoughts to myself. Their help was invaluable.

Derek Bell and Graham Long offered much instructive criticism on early versions of this work. I would also like to thank Bernadette Buckley and Branwen Gruffydd Jones, who inspired and encouraged me to continue my academic work. My uncle, Phil Chambers, also offered me much advice and inspiration during the formative stages of my studies. It was their encouragement that led me down this path.

My mum read through many of these chapters, offering comments on the spelling, grammar, and syntax. I owe her many thanks for this, along with her support throughout.

Instructive feedback on some of the work in this thesis was offered at the MANCEPT Workshops and the Geneva Graduate Conference in Political Philosophy. I would like to thank the organisers of these events, and all the participants.

Thanks are due to the Economic and Social Research Council who funded this project, and without whom this thesis would not have been possible.

The School of Social and Political Sciences at the University of Glasgow have provided a supportive environment throughout the course of my work on this thesis. I owe thanks to the Political Philosophy and Theory Group, and the Politics Seminars, from which I gained helpful feedback and advice. I would also like to thank all of the staff and my fellow postgraduate researchers.

Finally, many thanks are due to my friends and family for their support.

Author’s Declaration

“I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.”

Printed Name: Luke Armstrong

Signature: L.Armstrong

Chapter 1 – Introduction

1.1 Political Liberalism: Autonomy and Free Will

In his 1993 work *Political Liberalism*, John Rawls attempts to show how his theory of justice is not dependent on the truth of a particular doctrine within religion or philosophy. A reasonable person could accept this theory of justice whatever her other religious or philosophical commitments. In this thesis, I argue that Rawls' position cannot hold. Rawls considers people to be autonomous, and I show that his conception of autonomy contains metaphysical assumptions. These metaphysical assumptions reveal a flaw in Rawls' supposed neutrality between reasonable philosophical doctrines, and also a weakness in how he imagines a just liberal order can be established and sustained. Rawls considers people come to commit to his theory of justice through their own free will and intellectual resources. Data drawn from neuroscience and psychology, however, suggests that this approach is overly optimistic. This data challenges how we think about free will and our intellectual capacities. If we are to enable citizens to become autonomous – and for justice as fairness to be realised, citizens must be autonomous – a radically different approach must be taken from that proposed by Rawls.

Though Rawls argued that metaphysics had no place within politics, he does allow for the implications of scientific knowledge to be considered. Within what Rawls calls public reason – the method of deliberation by which conclusions regarding constitutional essentials are drawn – citizens can draw on science to support their reasoning. This gives us grounds, I argue, on which to support our political reasoning with the conclusions of neuroscience. There is work within neuroscience looking at whether we have free will. Considering such scientific knowledge leads us to revise how we think about autonomy.

A society without those Rawls describes as fully autonomous citizens – citizens motivated to act in accord with what is just – is an unstable one. Considering neuroscience may lead us to think that we cannot expect people to become fully autonomous if left to their own devices. Society should therefore be devised as to encourage the development of full autonomy. In this thesis, I argue that this requires us to adopt certain measures which are perfectionist, as they recognise a prior standard of the good against which to judge. Full autonomy provides such a standard; a person who is fully autonomous is living a better life than someone who is not. If the empirical data taken from neuroscience is true, then we should consider how important, yet vulnerable, autonomy is for the stability of a society. We would prioritise the development of autonomy over other political values.

1.2 Why Care About Autonomy in Political Liberalism?

There is a sense in which Rawls imagines that moral and intellectual powers underpinning autonomy are innate within the human condition. The basis of what he calls rational autonomy is in the two moral powers: the capacity for a sense of justice and the ability to formulate a conception of the good (Rawls 2005, pp.72-77). Because we have these two powers, we can be considered rationally autonomous when we develop them. The way in which we are autonomous politically, or fully autonomous as Rawls puts it, rests on how we develop the first moral power in relation to the society we inhabit (Rawls 2005, pp.77-81). Rawls recognises that we cannot be considered fully autonomous unless we live in a society in which the principles of justice are widely recognised and understood. The basis on which we come to realise the conditions necessary for full autonomy rests on our capacity for rational autonomy. That is, it is because we recognise our self-interests as rationally autonomous agents that we desire to live in a society in which we can become fully autonomous. Thus, the way in which we are supposed to become autonomous rests on certain assumptions regarding our innate sense of morality.

Due to these assumed characteristics, Rawls assumes that we will desire to become autonomous. Autonomous citizens are committed to upholding just institutions, in which liberal ideals are embedded. However, empirical evidence suggests that people are less committed to liberal ideals than liberals such as Rawls imagine. A study by the Hansard Society in 2019 found that 54% of the British public thought that a "strong leader who was willing to break the rules" was necessary. In the same survey, 42% thought that government would be more effective if it did not have to worry about votes in parliament. The Pew Research Center assessed attitudes towards democracy across 34 countries between 2018 and 2019, finding that 52% of people were dissatisfied with the way their democracy was functioning, while commitment to certain democratic ideals was found to be low (Wike & Schumacher 2020). Nationalist leaders have been gaining support across democracies in Europe, with some in positions of power satisfying the desire for a strong leader willing to break the rules. Responding to the Covid-19 crisis, Hungary's Victor Orbán passed a law allowing the government to rule by decree (Charlemagne 2020).

Though a rigorous assessment of attitudes towards democracy is not the focus of this thesis, these are signs that many people living in democracies are dissatisfied with the workings of democracy. What is of significance here is not the reasons why people are dissatisfied with democracy. There could be many legitimate reasons why a person became dissatisfied with the workings of democracy: corruption; politicians failing to listen to the people; a feeling that the political class is removed from the experiences of the rest of the population. Instead, the concern here is why people express dissatisfaction with democracy itself, rather than demanding the need for reform. This suggests a low level of commitment to the value of democracy on the part of many citizens. If many citizens possess a weak motivation to uphold just institutions, a problem is posed for the implementation of liberal political theory.

I argue that this problem is fundamental for Rawls. A Rawlsian might argue that if what Rawls refers to as "the basic structure of society" had functioned more effectively (Rawls 1971, pp.7-11), realising the sense of justice that Rawls argued should be embedded in this structure, then none of this would have been a problem. People would have continued to value liberal democratic institutions providing they functioned effectively, even without a deeper commitment to their importance on the part of many citizens. It is my contention in this thesis that this is mistaken. Whether or not the basic structure of society is functioning effectively, a citizenry that cares about the value of democracy and acting so as to realise just societies and communities is important. Without such a citizenry, the task of realising a just society will remain perpetually unstable.

This is part of the reason why I argue we should reassess the place of autonomy within political liberalism. Whereas the development of full autonomy is prescribed – a citizen should recognise the principles of justice and develop her sense of justice accordingly – the citizen is free to develop her rational autonomy within the bounds of what is reasonable. She must ensure that her conception of the good is compatible with a society which functions according to fair terms of cooperation. If this is the case, then her conception of the good is reasonable. As she formulates and acts to realise such a conception of the good, Rawls considers her rationally autonomous. The citizen is, then, left to her own devices to determine how she conceives of the good life, living life in accord with this conception. A person becomes fully autonomous when she comes to act from the principles of justice. Full autonomy, unlike rational autonomy, is not a matter of individual choice, but one of what would be collectively agreed in a hypothetical situation.

While the fully autonomous citizen is a citizen who acts from the principles of justice accepted by all other reasonable citizens, the rationally autonomous citizen is free to choose her good for herself.

Drawing on evidence from neuroscience and its philosophical implications, I argue that the conditions under which a person develops her rational autonomy can potentially undermine the development of full autonomy. Evidence within neuroscience suggests that our capacity for free choice is limited. Benjamin Libet precipitated much debate between philosophy and neuroscience on the question of free will (Libet et al 1983). As his experiments appeared to reveal that our conscious awareness of our intentions followed behind our brain's formulations of these intentions, Libet posited that our ability to consciously initiate our own intentions and desires to act was severely restricted. While philosophers have responded to Libet's work with a variety of convincing rebuttals, there are implications drawn from contemporary neuroscience that I argue remain problematic. We should, therefore, take the implications of neuroscience seriously. The determinism of neural processes and the lack of a centre of consciousness from where these processes are controlled should lead us to reassess the way in which we come to hold the thoughts we do. Whatever philosophical conception of free will one endorses, we should consider this problematic. If there is no centre of consciousness from where we can control our thought processes, then we also have little ability to control the way in which our thoughts are influenced. There is no centre of control from where we can regulate our thoughts. What we end up thinking and valuing is, then, largely outside of our control.

Without a prior commitment to the value of justice as fairness, we could not suppose citizens would continue acting as fully autonomous citizens. In a society in which ideas that threaten the stability of democratic institutions are prevalent, we could not expect a person to remain uninfluenced by these ideas. A person could be exposed to ideas on social media that were at odds with the values of political liberalism, deciding to reject political liberalism in favour of these ideas. In such a case, with no centre of conscious control from where thought processes are regulated, we could not expect this person to bring her ideals back into alignment with the values of political liberalism. She has little ability to determine how her surroundings influence how she conceives of the good. Without any prior commitment to political liberalism, she may well find alternative doctrines convincing.

Though Rawls downplays the role autonomy plays in political liberalism, claiming that it is not a regulative principle guiding other aspects of the theory (Rawls 2005, p.78), without fully autonomous citizens, none of the aims of political liberalism could be achieved. In a society in which the majority of citizens are not fully autonomous, there are few people left to uphold the principles of justice. The stability of the well-ordered society imagined by Rawls is lost. If we cannot assume that citizens will become fully autonomous through their own intellectual capacities, society should be structured so as to encourage the development of fully autonomy. While full autonomy, Rawls argues, is only applicable in the political domain – it is in realising the principles of justice politically that one becomes fully autonomous – I posit that we need to think about the value of full autonomy throughout our lives. This brings us close to what G.A. Cohen calls the "egalitarian ethos" (1992; 2008). In a society in which the egalitarian ethos is central, people refer to their concern for equality when justifying their actions. Likewise, the fully autonomous citizen recognises the importance of realising a sense of justice, but also recognises the way in which certain acts can impede the development of autonomy for others, in turn, eroding the sense of justice existing within society. Leaving a citizen to determine her own good may lead her to develop a moral psychology at odds with justice as fairness. Instead, the development of an appropriate moral psychology should be the focus of education. This is not to say that the state need determine all aspects of a citizens' good, but that the state should ensure its citizens are committed to upholding justice.

In *Eichmann in Jerusalem* (1963), Hannah Arendt focuses on the problem of people who have forgotten how to think and judge in totalitarian regimes. This problem should be taken seriously. It is not only a problem within totalitarian regimes. People fail to think and judge in liberal democracies, too. Arguably, this is part of the reason for the existence of what Rawls might have deemed "unreasonable" ideas in liberal democracies: the belief that an authoritarian leader who would break the rules would be preferable to democratic norms; the idea that we should judge people based on their race, gender, or sexuality; the belief that basic rights are not inalienable. Such ideas are unreasonable as they break the terms of fair cooperation. Parties in Rawls' original position – his thought experiment in which a position of impartiality is constructed in order to determine what is just – would not accept that such conclusions were just. However, these ideas persist within liberal democracies. Arendt argued that the failure to think and judge is most problematic in those who unquestioningly follow orders. Considering the problems revealed by neuroscience, we may conclude that there is no independent space from where we can question our ability to think and judge. Most of us unquestioningly accept many beliefs and practices as acceptable as we never have the opportunity to question their acceptability. It is not only the following of orders without question that leads to the problem of unreasonableness, but the lack of an ability to consciously control our thought processes independently of our environment. In an environment in which unreasonable ideas predominate, it could be expected that many people would become unreasonable. An individual under the sway of unreasonable ideas has little ability to control how these ideas influence her conception of the good. Therefore, rather than leaving individuals to determine their own good, the development of full autonomy should take precedence. Through a system of moral education, citizens should become committed to the value of fairness in the Rawlsian sense. The aim of this is to uphold the stability of the sense of justice within a society, but also to allow individuals to better realise their own autonomy.

1.3 Autonomy and The Principles of Justice

A question remains regarding how this argument relates to the central component of Rawls' theory: the principles of justice. Rawls presupposes that in a just society – one in which the principles of justice have been applied to the central institutions – citizens will become fully autonomous providing they are made aware of the duties placed on them (Rawls 2005, p.78). Knowing the content of the principles of justice due to its being embedded in society's institutions, citizens will come to act in accordance with the principles. I argue against this presumption. With nothing to motivate citizens to endorse the principles of justice other than knowledge of their content and effectiveness in practice, in a society in which freedom of conscience is protected, citizens may choose to endorse doctrines entirely at odds with justice as fairness. Knowledge in itself is not enough to commit a person to a doctrine.

If there is to be wide acceptance of the principles of justice, a prior commitment needs to be ensured to full autonomy. It cannot be assumed that citizens will act as fully autonomous citizens ought to. Thus, the relationship between the principles of justice and the autonomy of citizens needs to be reconsidered. In his formulation of the principles of justice, Rawls does not speak about autonomy. The only instance in which autonomy and the principles of justice are considered together is when Rawls states that a fully autonomous citizen acts from the principles of justice (Rawls 2005, p.77). However, here Rawls assumes that people will act as fully autonomous citizens should. The principles of justice are not devised so as to encourage the development of full autonomy, nor does Rawls assume the prior existence of citizens committed to full autonomy. The assumption that citizens will become fully autonomous is, then, based only on the notion that citizens will develop full autonomy through knowledge of the principles of justice. This is, I argue, a weak claim on which to ground the stability of justice as fairness.

In a society in which there were no fully autonomous citizens, no one would act in accordance with the principles of justice. Thus, in such a society, the realisation of justice as fairness would be impossible. While Rawls states that full autonomy develops through citizens coming to understand justice as they come to recognise the basic structure as just, without fully autonomous citizens to uphold a just basic structure, we are left with no one to ensure the basic structure is just. Within the principles of justice is, then, an implicit assumption that citizens are fully autonomous. As the existence of fully autonomous citizens is essential for political liberalism as Rawls formulates it, I argue against making this assumption. Instead, a commitment to the development of full autonomy is needed if citizens are to act from the principles of justice; this commitment needs to be fostered.

The application of the principles of justice must, then, be rather different from how Rawls imagines. If the priority of liberty is granted to the institutions constituting the basic structure of society, with little regard given to how individuals will behave within this structure, individuals may use the liberty granted to the detriment of the principles. This is why, following Cohen (1992; 2008), I argue that the behaviour of individuals must be considered. Without a citizenry committed to upholding the principles of justice, the stability of the well-ordered society for which Rawls argues is lost. Therefore, the principles of justice must not only be applied across society, but also in regard to the behaviour of individuals. It is this line of reasoning that undergirds my argument for the need of moral education. If the ideals Rawls formulates are to be attained, citizens must become fully autonomous; I argue this can only occur through moral education.

1.4 Autonomy in Political Liberalism: An Overview

In Chapters 2 and 3, I explain the basic argument of political liberalism and how autonomy is formulated within this theory. Chapter 2 begins with an assessment of the main concepts developed by Rawls in *A Theory of Justice*. Rawls' central task here is to define justice. He does this through what he calls the original position and the veil of ignorance (Rawls

1971, pp.136-142). Parties, as he calls those within the original position representing citizens, are to imagine they are behind a veil of ignorance; they are ignorant of citizens' place in history, their position within society, and many other aspects of their lives which could be considered controversial. The aim here is to construct a position of impartiality. The parties will be impartial as they do not know where in society citizens will come to exist; without knowing this, during deliberations they cannot knowingly favour certain principles which will favour those citizens over others. From this position of impartiality, Rawls explains the reasoning that leads towards the acceptance of the two principles of justice. These principles are representative of the theory of justice as fairness.

Political liberalism is then considered, and the ways in which Rawls developed the concepts and methods within *A Theory of Justice*, in order to show that they do not presuppose the truth of one philosophical or religious doctrine. He does this by explaining the appropriate ways of reaching political consensus. When deliberating over constitutional essentials, we should respect public reason (Rawls 2005, pp.213-216). Public reasons are reasons acceptable to people regardless of their other philosophical or religious commitments. Justice as fairness is one such concept. People can accept justice as fairness whatever their beliefs. This then forms the basis for the overlapping consensus of reasonable doctrines (Rawls 2005, pp.144-150). Society, for Rawls, is based on reasonable people united in a sense of justice, but who differ in their beliefs otherwise. People's beliefs overlap where the sense of justice is shared.

Rawls identifies two forms of autonomy in political liberalism: rational and full. The rationally autonomous citizen is able to determine her own good, living in accord with this determination, and taking responsibility for the consequences arising. The fully autonomous citizen acts from the principles of justice chosen in the original position. Whereas rational autonomy relates more to the form of autonomy important to Raz (1986) and Mill (1859) – for whom autonomy is the capacity for self-realisation – full autonomy is akin to Kantian autonomy; a form of moral autonomy on which an individual determines how to act by reflecting on the principles of practical reason. For Kant, this means identifying the categorical imperative: acting as though the act would become a universal law (Kant 1797). Likewise, Rawls' fully autonomous citizen reflects on what would be determined as just from the perspective of the original position. Whereas rational autonomy – or autonomy as self-realisation – provides the central concern of political theory for Raz and Mill, and moral autonomy regulates all aspects of political and moral philosophy for Kant, autonomy is not Rawls' central concern. Instead, autonomy is a by-product of other political considerations. With the right conditions in place, Rawls assumes that citizens will act as rationally and fully autonomous agents. In a just society, individuals will be able to determine their own good and act from just principles, fulfilling the demands of rational and full autonomy.

Though Rawls did not place autonomy at the centre of his political theory, as Mill and Kant did, autonomy is more important for his theory than Rawls imagines. This is particularly the case with full autonomy. Without fully autonomous citizens, the realisation of justice as fairness would be impossible. Any political doctrine requires people who are convinced of its value, and motivated to act in accord with it. If the majority of citizens in a society are not concerned with the requirements of justice, and fail to act accordingly, justice will be unrealisable. The form of justice devised by Rawls requires citizens act in certain ways. In order to realise the principles of justice, citizens need to agree on a certain constitutional form. They need to vote for certain policies, and ensure that society is structured as to promote justice. This requires a citizenry possessing a certain form of moral psychology. As Rawls deems citizens rationally autonomous, he requires them to choose their own good. Determining their own good, citizens will not necessarily develop the kind of moral psychology necessary to uphold political liberalism.

In Chapter 4, I explain the methodology to be used in subsequent chapters, and my approach to the relationship between theory and empirical problems. Rawls' own method of reflective equilibrium is adopted. In reflective equilibrium, judgements and principles are assessed against one another; where there is dissonance between the two, adjustments are made to bring both into agreement (Rawls 1971, pp.48-51). Rawls later distinguished "wide" from "narrow" reflective equilibrium (Rawls 1974). While in narrow reflective equilibrium an agent only needs to make judgements in "conviction and confidence", in wide reflective equilibrium, "certain conditions of rationality" are satisfied. To satisfy these conditions, an agent needs to have considered all plausible alternatives to reach a state where she is confident in the judgement. Norman Daniels argues that within wide reflective equilibrium, not only judgements and principles need to be considered, but also what he calls "background theories" (Daniels 1979). Though Daniels does not explicitly state what theories can be categorised as background theories, it can be surmised that various scientific and philosophical theories can be included within reflective equilibrium, which are then assessed against the principles and judgements. I argue that the conditions of wide reflective equilibrium should hold across the framework of political liberalism, and neuroscience should be considered at all stages. To do so would, I argue, alter the conclusions reached. I establish the grounds on which we should consider science and its relationship with politics in this chapter.

Thus, rather than assume the nature of a person's moral psychology, we should, I argue, test any theory of moral psychology against the empirics. I argue that Rawls' allowance for the use of scientific theories to support public reasons should enable us to consider psychology and neuroscience. However, the close relationship between science and metaphysics should, I argue, lead us to consider the metaphysical implications of science, where these implications bear on matters within political theory.

In Chapter 5, I explain the main metaphysical theories of free will, before arguing that Rawls requires the truth of free will – whether that of the compatibilist or the libertarian – for the purposes of political liberalism. Rawls requires citizens who possess the moral and intellectual capacities to regulate their own thought processes, ensuring they cohere with the ideals of political liberalism. If citizens lack these capacities, then political liberalism is undermined. However, that they are left to their own devices to develop these capacities suggests that Rawls considers them to possess some form of free will. If it were argued that hard determinism was true, and thus we could not be considered morally responsible, Rawls would have to reject this view. Such a view undermines the basis on which Rawls thinks about people, their lives, and their abilities to make agreements with one another in order to live sociably. Rejecting the basis on which people come to make these agreements would require a different way of thinking about human nature and moral psychology.

Arguments concerning free will in neuroscience are assessed in Chapter 6, before the philosophical implications of these arguments are more fully explained in Chapter 7. Considering the implications of neuroscience should lead us to revise the way we think about people, their thoughts and actions. Though the compatibilist perspective on free will is somewhat immune to the implications of empirical evidence, the data within neuroscience should lead us to revise how we think about thought processes. If we have little conscious control over our thought processes, then there may be little we can do to prevent the influence of unreasonable ideas. While compatibilist arguments hold in securing responsibility for our acts – nothing taken from the neuroscience undermines the arguments of Harry Frankfurt (1971), for instance – there remain problems unresolved in relation to our thoughts. There are two problematic implications here. First, the conscious will is not a prime mover (Haggard 2008). Instead, neural processes occur in deterministic cycles. While this in itself is not fatal to the compatibilist position – a compatibilist would expect the will to be determined by prior causes – the second implication proves more troubling. This is that there is no centre of consciousness from where these processes are controlled (Roth 2003). Whether consistency holds between two thoughts is, then, essentially a matter of luck; that is, thought processes are subject to little human control. Without a centre of conscious control from where thoughts can be regulated, we could not expect someone to necessarily develop the necessary moral psychology for political liberalism if her thoughts were the result of unreasonable external influences.

Assessing Rawls' formulation of autonomy against ways of thinking about free will within neuroscience, in Chapter 8, I argue that we should revise how we think about autonomy. However, neither rational nor full autonomy are entirely irreconcilable with the data in neuroscience. Whether or not we have free will, we may be capable of forming a reasonable conception of the good and acting justly. However, rather than assume there is an innate sense of morality within the person – as Rawls does when he claims we have two moral powers – we should recognise that any sense of morality we come to hold is dependent on neural processes over which we hold little influence. Rather than assign individuals the responsibility of developing their sense of morality, we should think of a societal responsibility to shape the ideas which influence our moral development. Considering this from the perspective of rationally autonomous parties in the original position – for whom, as this is a hypothetical space, the data on neural determinism does not apply – it will be recognised that without fully autonomous citizens, justice as fairness will be unrealisable. Full autonomy cannot be considered a by-product of the social structure; it may be the case that even with a broadly just social structure in place, people are still persuaded by unreasonable ideas. The development of full autonomy should therefore be prioritised. Assuming that citizens are rationally autonomous in the appropriate way may undermine full autonomy. Citizens may develop their rational autonomy in ways that are at odds with the values of justice as fairness. If we take the implications of neural determinism seriously, then we cannot expect citizens to ensure the thoughts they come to think are reasonable, and that they come to value the appropriate sense of justice. Instead, society should be structured as to encourage people to develop their full autonomy.

Therefore, I argue in Chapter 9 that if justice as fairness is to be realisable, aspects of perfectionism are necessary; citizens should be expected to be fully autonomous. The fully autonomous citizen is a necessary component of justice as fairness, and the person living a fully autonomous life is living a better life than one who is not, according to the standards of political liberalism. Following from this are several practical implications. A system of moral education is required. The purpose of education should be to inspire children to recognise the importance of fairness. As adults, people should be motivated to act as to realise justice as fairness. We cannot assume that they will do so of their own free will. The effect of this is likely to reduce the diversity of beliefs within a society. For instance, a person raised to value fairness in the Rawlsian sense is unlikely to endorse political libertarianism. Stability, on the other hand, will be enhanced through a common understanding of fairness and its value.

As referenced earlier, evidence suggests that citizens are currently detached from the value of liberal democracy. This problem could perhaps be resolved through the perfectionist approach suggested here. On this revision of political liberalism, rather than leave citizens to determine their own good, citizens should be motivated to become fully autonomous, viewing this as part of their good. Acting in accord with full autonomy across their lives, they act justly towards each other, and recognise the value of the institutions that promote this sense of justice. Such citizens would be motivated to uphold the ideals of a liberal democracy.

Chapter 2 – The Basic Argument of Political Liberalism

2.1 Introduction

In *Political Liberalism*, John Rawls attempted to formulate a theory of justice that would be compatible with the plurality of philosophical, religious, and moral beliefs existing in modern societies. This chapter explains the basic argument of *Political Liberalism*, expanding on Rawls' key ideas and how they fit into the scheme of political liberalism. Ideas that specifically relate to Rawls' conception of the person and autonomy will be explained fully in the following chapters. Thus, they will only be briefly discussed here. The purpose of this chapter is to detail the overall structure of political liberalism, so that the role of autonomy can be explained at a later stage. When Rawls' theory of autonomy is critiqued in subsequent chapters, it will then be clearer how the aspects of autonomy being critiqued relate to Rawls' overarching argument. Finally, some of the main critiques of political liberalism are discussed at the end of this chapter. These will be returned to throughout the following chapters, as I assess whether they are supported or undermined by the implications arising from this critique.

In 2.2, the theoretical content of *A Theory of Justice* is examined, and I explain its central ideas, which are integral to Rawls' general thought. This is followed by an explanation of the two principles of justice in 2.3. Across these two sections, justice as fairness – the central idea within Rawls' political thought – is defined. A discussion of Rawls' understanding of publicity and public reason is undertaken in 2.4. These ideas show how Rawls imagines citizens come to endorse justice as fairness, and are key to its stability. Public reason also establishes the limits of political deliberation; if citizens are to accept the legitimacy of state power, then the arguments undergirding this legitimacy must be acceptable to them. *Political Liberalism* is then analysed in 2.5. Justice as fairness in Rawls' later work becomes a political conception of justice, supposedly detached from wider philosophical considerations. The way in which Rawls formulates this idea is explained in 2.6, through the idea of the overlapping consensus. Finally, critiques of *Political Liberalism* are assessed in 2.7. In later chapters, I return to these critiques, showing how they relate to the critique established throughout this thesis.

2.2 A Theory of Justice – The Theory

This section explains the key ideas of Rawls' theoretical constructs, such as the basic structure, the original position, the veil of ignorance, primary goods, and goodness as rationality. These ideas show how Rawls draws out his argument, and provide support, for his conception of justice. For Rawls, justice is synonymous with fairness. While there are changes to the formulations of this argument throughout Rawls' work, the overall argument for justice as fairness remains the same.

In formulating a theory of justice, Rawls rejected utilitarianism and intuitionism, and instead returned to the idea of the social contract (Rawls 1971, VIII). Kantian ideas were restructured into a framework that Rawls intended to strengthen the idea of justice. According to Rawls, justice is the "first virtue of social institutions" (Rawls 1971, p.3). For a well-ordered society to function, its institutions must be just. Furthermore, the idea of justice undergirding these institutions must be one that is generally agreed upon, otherwise the lack of consensus will cause ruptures within society. A society with a generally agreed upon conception of justice at its heart would allow for the plans of individuals to cohere, enabling all to pursue their own aims within the framework of a well-ordered society (Rawls 1971, p.6).

According to Rawls, "the primary subject of justice is the basic structure of society" (Rawls 1971, p.7). The basic structure consists of all the major institutions across society, and legal protections such as freedom of thought, free markets, private property, and the family. These institutions have a major effect on the lives of a society's inhabitants. They decide the rights and duties people have, they determine peoples' aspirations, and they shape the types of lives people lead. For a society to be just, its basic structure must be just.

The task for Rawls was to identify this conception of justice. Rawls did this by appealing to the idea of the original position (Rawls 1971, pp.17-22), an initial situation in which all are equal, assumed to be rational, and able to agree upon principles of justice through a process of deliberation. It is a purely hypothetical situation; Rawls did not assume such a place exists. Furthermore, when Rawls refers to parties in the original position, he is referring to representatives of citizens in the well-ordered society, not the citizens themselves (Rawls 1971, p.64). Along with attempting to establish the conditions necessary for impartiality, this is to avoid committing to any particular philosophical conception of the self that not all would necessarily agree was true. Parties in the original position are placed behind a veil of ignorance (Rawls 1971, pp.136-142). This veil ensures that the parties have no access to certain forms of knowledge: their place in society, their share of society's resources, their conception of the good, their way of life, particular psychological features, or their own natural talents and abilities. Moreover, they are unaware of the general level of development of the society in which they exist. From this position, they must determine principles of justice, and accept that they will live their lives according to the implications of those principles. Rawls argues that the securing of primary goods will be at the centre of deliberations (Rawls 1971, pp.91-95). Whatever needs and desires people may have that are personal to them, there are certain primary goods "that every rational man is presumed to want" (Rawls 1971, p.62). There are goods that it is in everyone's self-interest to want, to satisfy their biological needs and meet the demands of their basic moral psychology. Primary goods are divided into social and natural goods. Primary social goods are, for example, "rights and liberties, powers and opportunities, income and wealth" (Rawls 1971, p.62). Primary natural goods are goods such as "health and vigor, intelligence and imagination" (Rawls 1971, p.62).

According to Rawls, self-respect is the most important primary good (Rawls 1971, p.440), as it is through having self-respect that individuals are motivated to pursue their own ends; other primary goods enable a person's self-respect to be fulfilled. At the basis of Rawls' understanding of self-respect is the assumption of what he calls the Aristotelian Principle. This is the idea that individuals enjoy exercising their "realised capacities" (Rawls 1971, p.426). All of us have certain natural talents. As we develop, we come to master these talents, finding satisfaction in more complex exercises of these talents the further our talents grow. Self-respect, then, is witnessed in those who are successful in realising these capacities (Rawls 1971, p.440). When we feel satisfied that we are reasonably successful in life due to the realisation of these capacities, we come to respect ourselves. For Rawls, self-respect grounds the basic motivation of individuals. It is because of their self-respect that they pursue their own ends, finding fulfilment in achieving their aims in life. This virtuous circle also supports the stability of a society, as the continuous reinforcement of social cooperation over time motivates individuals to act in accordance with the principles of justice. Self-respect is of vital importance for Rawls. it also forms the basis of Rawls' conception of moral psychology in *Political Liberalism*, (Rawls 2005, pp.81-82) on which citizens are viewed as desiring to be fully functioning members of a cooperative society. This will be further discussed in 2.3.

Where Rawls speaks of ‘the good’ in relation to primary goods, he has in mind the “thin theory of the good” (Rawls 1971, pp.395-399). This is not inclusive of deeper aspects of the good belonging to any one philosophical or religious tradition. For instance, primary goods are not goods because God deems them necessary for a satisfactory life, or because they are requirements for a life lived according to Aristotelian virtue. Instead, the thin theory of the good defines what will be central to all rational persons’ wants and needs, regardless of their philosophical and religious convictions. This leads Rawls to say that goodness is rationality. Goodness is what is desirable by rational persons. However, for Rawls, the right is prior to the good. We must determine our conceptions of the good within the confines with what is agreed to be right. This means that any theory of the good must be compatible with the principles of justice.

When seeking to define the principles of justice from the original position, the primary goods define the nature of what will be sought. Being rational agents, Rawls thought that each person would seek to maximise their own holdings (Rawls 1971, pp.118-119). However, other individuals would not accept conditions whereby they were disadvantaged by another’s gains, meaning compromise would be necessary. This compromise is reflected in the two principles of justice. According to these principles, we are afforded the opportunity to maximise our gains as is compatible with the same opportunity for others. The principles are further explained in what follows.

To conclude, the aim of the theory here is to identify a conception of justice that can act as a regulative principle for a society. Rawls achieves this by imagining what would be sought by parties in the original position, placed behind a veil of ignorance. Such parties seek the best conditions possible for those they represent. Compromise is sought between the parties, as to maximise gains insofar as maximisation is possible with the same level of maximisation for others. It is primary goods that parties are concerned with maximising. These goods are what Rawls deems necessary for a satisfactory life, the most important of which is self-respect. A person who has self-respect wishes to develop her moral sensibilities and the way of life she desires to live as a result, feeling satisfaction when she is successful in this task. To secure these goods, Rawls argues that a certain arrangement of rights and liberties is necessary.

2.3 A Theory of Justice – The Principles

Though they are revised throughout *A Theory of Justice*, and in later work, the original formulation of Rawls' principles of justice read as follows:

> First: each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others.
> Second: social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all (Rawls 1971, p.60).

The two principles of justice represent terms that would be acceptable to everyone. Securing liberties such as political liberty, freedom of speech, freedom of thought, freedom of assembly, the right to personal property, and freedom from arbitrary arrest, the first principle ensures that all individuals are able to advance their own aims in life without restriction (Rawls 1971, p.61). The second principle ensures that economic inequalities do not impede their ability to pursue their own aims, as any inequalities that do exist are to their advantage, as specified by the difference principle. By accepting the two principles, justice comes to be recognised as fairness.

Rawls thought that these principles would be agreeable to all rational agents in the original position. However, Rawls found two problems with the principles of justice: a problem of ordering, and a problem of interpretation. I explain the problem of ordering, before moving on to the problem of interpretation.

While determining the principles of justice, Rawls saw that there would be the problem of which principles took priority (Rawls 1971, pp.40-41). Rawls introduced the idea of lexical priority, the idea that principles need to be arranged in order of their importance (Rawls 1971, pp.42-43). Before a principle can be satisfied, the principles prior to it must be satisfied. Because self-respect is the most important primary good, the liberty to secure self-respect must take priority over all other rights and liberties (Rawls 1971, pp.543-545). Therefore, in Rawls' sequence of lexical priority, the first principle – equal liberties – comes before the second principle – social and economic inequalities. Social and economic distributions must be arranged as to be compatible with a system of equal liberties for all. While social and economic equality are important, there would be little point in attempting to secure equality in a society that lacked self-respect, and the stability of society that self-respect creates. People who lacked self-respect would not be motivated to secure either of the principles. Thus, the conditions for establishing self-respect must be secured before other conditions can be considered. This arrangement would be agreed on in the original position, as it would be recognised that certain rights and liberties were necessary for securing other rights and liberties (Rawls 1971, p.45). Lexical priority is of importance, as it is used throughout Rawls' work in order to determine issues of comparative importance.

I now turn to the problem of interpretation of the second principle. Noting that the terms "everyone's advantage" and "equally open to all" were ambiguous, Rawls suggested four possible interpretations: *a system of natural liberty, natural aristocracy, liberal equality,* and *democratic equality* (Rawls, 1971, p.65). Of these four interpretations, Rawls thought *democratic equality* was preferable.

The following table is used to explain the four interpretations:

	"Everyone's Advantage"	
"Equally Open"	Principle of Efficiency	Difference Principle
Equality of Careers Open to Talents	System of Natural Liberty	Natural Aristocracy
Equality as Equality of Fair Opportunity	Liberal Equality	Democratic Equality

(Rawls 1971, p.65)

As can be seen from the table, Rawls finds two ways of interpreting each phrase – "everyone's advantage" and "equally open" – leading to the four possible interpretations. I explain *the principle of efficiency* to begin. I then explain *equality of careers open to talents* (formal equality), and *equality as equality of fair opportunity* (fair equality), assessing how these principles lead to the *system of natural liberty* and *liberal equality* interpretations respectively. Following this, I explain the difference principle, and discuss how formal equality and fair equality lead to the *natural aristocracy* and *democratic equality* interpretations. Each interpretation will be defined in order to determine why Rawls views *democratic equality* as preferable.

Rawls draws the principle of efficiency from what economists term Pareto optimality (Rawls 1971, p.66). The principle of efficiency means that if a distribution cannot be altered in one person's favour without disadvantaging another individual, it is an efficient distribution (Rawls 1971, p.67). A *system of natural liberty* ensures that positions are formally open to all who are willing to work to attain them (Rawls 1971, p.66). Under a system of natural liberty, if the given distribution is efficient, and there is formal equality of opportunity, then the distribution can be considered fair. Formal equality of opportunity means that a position is open to all, and that individuals are not discriminated against. However, if an individual cannot afford the education necessary to attain the position, nothing need be done to rectify this situation if the aim is to achieve formal equality. With formal equality, a position need only be nominally open to all. *Liberal equality* is instead based on the premise of fair equality. This requires that positions are not only formally open to all, but that each person has a fair chance of attaining a position. If two people have equal talents and abilities, then they should have an equal opportunity of attaining the same position, regardless of economic and social factors. Rawls notes that according to the interpretation of liberal equality (Rawls 1971, pp.72-73), formal equality under a system of natural liberty would lead toward inequalities created by natural and social factors that have developed historically. Thus, some will be unjustly favoured or disfavoured, due only to whether their natural and social circumstances offered them the chance to develop their talents. Through liberal equality, an attempt to adjust this situation is introduced through the principle of fair equality of opportunity. Yet, with the principle of efficiency combined with fair equality of opportunity, there will remain the possibility of large social and economic inequalities that are influenced by arbitrary factors. These inequalities may prohibit some individuals from fully realising their talents and abilities (Rawls 1971, pp.74-75).

This leads Rawls to the alternative of the difference principle. The difference principle holds that inequalities are not to exist unless they are to the advantage of those with least (Rawls 1971, p.75). With formal equality of opportunity, this leads to the interpretation of *natural aristocracy*. In a system of natural aristocracy, the talents that people hold can only be used if their use is to the advantage of the less talented (Rawls 1971, p.74). As there is only formal equality of opportunity, however, all individuals do not have a fair chance of attaining positions or developing their talents. As Rawls held that unequal distributions influenced by arbitrary factors were unjust, Rawls claimed natural aristocracy was unjust. Although unequal talents are being used to benefit the disadvantaged in a natural aristocracy, there remains only formal equality of opportunity. All citizens do not necessarily have a fair chance of attaining positions. Rawls argued that the interpretation of *democratic equality* achieved the conditions necessary for these fair chances. With democratic equality, the difference principle is combined with fair equality of opportunity. Inequalities can only exist when they are to the advantage of those with less, and positions are attainable by all. Economic factors are not prohibitive as efforts are made to ensure that the disadvantaged have an equal chance of obtaining a position. Natural inequalities that exist only through historical chance are thus accounted for, and inequalities may only exist when they are to the advantage of all (Rawls 1971, p.75).

Through the two principles of justice, with the basic liberties being prioritised lexically, and the second principle being interpreted through democratic equality, with fair equality of opportunity taking priority over the difference principle, Rawls establishes what he calls justice as fairness. Free and equal persons choosing the principles of justice to undergird their constitution would choose principles that allowed fair terms of cooperation. These principles benefit all, allowing each person to realise her conception of the good (Rawls 1971, pp.11-17). This defines the idea of justice as fairness.

2.4 Public Reason and Publicity

Rawls imagines that in a society in which justice as fairness is realised, the publicity condition will be satisfied. That is, citizens will be aware of the requirements of justice, and will be motivated to act in accord. In *Political Liberalism*, Rawls then develops the concept of public reason; reasons that all citizens can endorse, regardless of their beliefs. In this section, I explain each concept, beginning with publicity.

> When the basic structure of society is publicly known to satisfy its principles for an extended period of time, those subject to these arrangements tend to develop a desire to act in accordance with these principles and to do their part in institutions which exemplify them (Rawls 1971, p.177).

Publicity is the idea that the general principles agreed on in a society are publicly known and accepted. Through the idea of publicity, Rawls argued that stability would be achieved. People recognise that they benefit from social cooperation (Rawls 1971, pp.176-

177). Due to this, they are motivated to cooperate with one another. The principles of justice lead naturally toward such a cooperative society. When people act according to the principles of justice, society becomes a "cooperative venture for mutual advantage" (Rawls 1971, p.4). Rawls views the relationship between cooperation and self-respect as a sort of virtuous circle: due to their self-respect, people are motivated to pursue their own ends; social cooperation assists people in pursuing their ends; their self-respect is then enhanced through social cooperation. When the principles of justice become publicly recognised, the self-respect of a society's members is increased, because it is recognised that the principles of justice assist individuals to advance their own ends through a system of social cooperation (Rawls 1971, p.178). This reinforces peoples' desire to act in accordance with the principles of justice, and, therefore, the principles of justice "generate their own support" (Rawls 1971, p.177). Rawls' argument here is based on a certain understanding of moral psychology. It is assumed that when people make agreements, they keep to them. They "love, cherish, and support whatever affirms their own good" (Rawls 1971, p.177.). As their own self-interest is both in accord with, and advanced by, the principles of justice, Rawls argued that there would be a strong motivation for individuals to accept and act according to the principles.

Publicity enhances the self-respect of individuals (Rawls 1971, p.178), which, as previously mentioned, Rawls thought was the most important primary good (Rawls 1971, p.62). When people act on their conceptions of the good within the limits of justice, subsequently realising a level of success, others come to respect them (Rawls 1971, pp.178-179). The respect of others enhances a person's own sense of self-respect. Thus, when the principles of justice are publicly known, and recognised as effective, the opportunities for all to enhance their self-respect are increased. People being able to realise their own aims in life due to the cooperative nature of the society in which they live become more self-respecting. This further increases their willingness to act cooperatively. Self-respect and cooperation are, then, mutually reinforcing.

Though the publicity condition still applies within political liberalism, the concept of public reason is developed in addition. Before explaining public reason, it will be helpful to understand what Rawls means by 'reason' and 'reasonable'. Rawls states that reason is "an intellectual and moral power, rooted in the capacities of (a society's) members" (Rawls 2005, p.213). Reason allows the members of a society to guide the decisions of that society, to make plans, and to prioritise their ends (Rawls 2005, pp.212-213). A reasonable person is distinguished from a rational person. Whereas rational people pursue their own self-interest, reasonable persons ensure that their self-interest is compatible with the interests of others, and of society as a whole (Rawls 2005, pp.48-51). Reasonable persons desire a "social world in which they, as free and equal, can cooperate with others" (Rawls 2005, p.50). There are two aspects of the reasonable: willingness to propose fair terms of cooperation and recognition of the burdens of judgement (Rawls 2005, pp.54-66). First, a willingness to propose fair terms of cooperation means that a person desires to enter into any social relations with others on terms other people will be willing to accept. This is key to understanding the difference between the reasonable and the rational for Rawls. A rational person knows her self-interests and is able to pursue them in social relations, but without being reasonable, a rational person would not necessarily propose fair terms of cooperation. Second, the burdens of judgement are the sources of reasonable disagreement (Rawls 2005, pp.55-58). Though a doctrine may be reasonable, it may nevertheless exist in tension with other reasonable doctrines due to fundamental disagreements, whether these disagreements be due to religious, metaphysical, empirical, moral, or social factors. In public reason, the burdens of judgement must be borne in mind, as there are certain issues that reasonable persons may never agree upon. Consideration of the burdens of judgement is the basis of toleration for Rawls. Reasonable citizens come to hold reasonable comprehensive doctrines, systematic modes of thought containing religious, philosophical, or moral ideas informing human life (Rawls, 2005, pp.58-66). A comprehensive doctrine is reasonable if, when taking into consideration the burdens of judgement, it is compatible with a socially cooperative world, and can be justified to other reasonable persons as being congruent with such a world. The burdens of judgement place a limit on the kinds of comprehensive doctrines that are permissible within a society. They also impose duties on individuals to ensure that their doctrines are reasonable, and furthermore, that their doctrines do not mandate the use of state power to suppress other doctrines (Rawls 2005, pp.59-61). These two aspects of the reasonable inform the idea of public reason.

There are two circumstances in which citizens exercise public reason: when matters of justice are being decided and in deliberation on the good of society (Rawls 2005, p.213). This occurs primarily in forums in which what Rawls calls "constitutional essentials" are being decided, or in matters of basic justice (Rawls 2005, p.214). Constitutional essentials are of two kinds: the structure of government, and the powers of each branch; basic rights and liberties and who they apply to (Rawls 2005, p.227). Either of these matters must be determined within the limits of public reason. Public reasons are reasons that all people can accept. Thus, whatever my own beliefs in religious or philosophical matters, I can accept the content of public reason. When settling constitutional essentials, we should not, then, appeal to the wider aspects of our religious or philosophical beliefs, as all people might not accept such reasons (Rawls 2005, pp.224-225). If I was to appeal to my religious faith in order to argue that all branches of government should be subordinate to the church, many people would not accept my reasons. This is because my reasons are not public reasons. To devise a constitution that is acceptable to all citizens as moral equals, the reasons we use to support constitutional essentials must remain strictly within the political domain. This is what makes a reason a public reason.

Though forums in which constitutional essentials are decided are the primary subject of public reason – Rawls writes that the supreme court is the "exemplar of public reason" (2005, p.231) - it is preferable, Rawls claims, if we stay within the limits of public reason throughout political deliberations (Rawls 2005, pp.215-216). These forums are considered to have a primary duty to respect the limits of public reason because if they are not respected there, they will be respected nowhere. Politicians should keep their campaigns within the limits of public reason, and citizens should respect public reason when deciding how to vote if constitutional essentials are being considered. Reasonable citizens – citizens whose conceptions of the good cohere with those of others in a cooperative society – realise the importance of reasons used to support political principles being agreeable to all other reasonable citizens. If state power can be used to prevent certain activities from being performed, then it is necessary that there is common agreement on what is the legitimate use of this power. To reach this agreement, the bounds of public reason must be respected.

To sum up, if citizens are to recognise the value of justice as fairness, the publicity condition must be reached. The political principles undergirding the constitutions must be recognisable in the institutions forming the basic structure of society. Citizens are then aware of their equal basic rights and liberties, and the way in which the arrangement of society is conducive to their own self-interests. From this, a citizen's sense of self-respect is increased. She realises that her aims in life are supported through the arrangement of the society in which she lives. She is afforded the opportunity to realise her own plan of life, derived from the conception of the good she has conceived. When she is successful in realising these plans, she is respected by others, and increases her own self-respect in turn.

She is then motivated to act cooperatively; the benefits of cooperation and self-respect exist in a virtuous circle. Rawls then develops the concept of public reason in *Political Liberalism*. An understanding of public reason helps to explain the solution Rawls finds to the central problem of political liberalism: how a conception of justice can be endorsed by people regardless of their philosophical and religious commitments. A public reason is a reason acceptable to all other citizens. Public reasons are void of commitments to religious and philosophical positions. Whatever a person's beliefs, if she is reasonable, she should accept the public reasons used to support the adoption of political principles. In part, this answers the problem of political liberalism.

2.5 Political Liberalism: Revising Goodness as Rationality

I now turn to how Rawls revised his theory of justice in *Political Liberalism*. To begin, I explain how Rawls clarifies the line of reasoning from the original position to the acceptance of the principles of justice. I then turn to how Rawls revises the thin theory of the good to show how it is not a comprehensive theory of the good. These ideas identify what is being sought through formulating a theory of justice. To address the way in which a society could support this theory of justice, Rawls further develops the idea of the basic structure in *Political Liberalism*, which is supported by the basic liberties. After examining these ideas, I discuss how Rawls further refines the idea of justice as fairness. With the content of justice and its implementation explained, I explain the specific problem Rawls was responding to in *Political Liberalism*: the issue of how a conception of justice could be endorsed in a society divided by a range of comprehensive doctrines.

In *Political Liberalism*, Rawls further refined and developed ideas originating in *A Theory of Justice*, responding to its critics. The main problem Rawls sought to address was whether justice as fairness was a comprehensive conception of justice (Rawls 2005, pp.XVI-XVII). If it were comprehensive, citizens endorsing such a conception would be committing to the various metaphysical and moral doctrines to which it was tied. The problem here is that a modern democratic society that allowed for freedom of thought would, over time, develop a range of comprehensive doctrines. Its citizens would be divided by various religious and philosophical beliefs. These comprehensive doctrines would potentially be in discord with the comprehensive doctrine of justice as fairness. The legitimacy of the democratic regime would, then, be lost amidst the range of comprehensive philosophical doctrines. Rawls attempted to refine the conception of justice toward a specifically political conception of justice, one that avoided moral or metaphysical commitments. Citizens would be able to endorse the political conception of justice without rejecting their own comprehensive doctrines.

In the line of reasoning that leads to the acceptance of the principles of justice, Rawls stresses that no metaphysical positions are presupposed. Rawls aims here to counter those such as Michael Sandel who argue that the original position assumes an artificial account of human nature (Sandel 1984). No such perspective on human nature is assumed in the original position, according to Rawls (2005, p.27). The original position is a hypothetical situation. We do not need to assume anything about human nature in order to follow the line of reasoning Rawls devises with the original position. Instead, all we need to do is imagine parties existing in the original position who understand we will have certain interests to satisfy. From this point of departure, Rawls then argues towards the principles of justice.

Rawls developed the two principles of justice across *A Theory of Justice* itself, and in subsequent work, taking up a revised version of the principles for political liberalism, which read as follows:

a. Each person has an equal claim to a fully adequate scheme of basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.
b. Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least advantaged members of society (Rawls 2005, pp.5-6).

The task here, however, is less to explicate justice, but to answer Rawls' question: "How is it possible that deeply opposed though reasonable comprehensive doctrines may live together and all affirm the political conception of a constitutional regime?" (Rawls 2005, p.XVIII). To achieve this, Rawls attempted to define a political conception of justice that could be endorsed within an overlapping consensus of reasonable comprehensive doctrines (Rawls 2005, p.10). The political conception of justice would legitimise the democratic regime itself, providing the basis for a stable and well-ordered society. However, before explaining how the conception of justice is purely political, it will be necessary to briefly explore how Rawls thinks about justice in *Political Liberalism*.

The aim of *Political Liberalism* remains to find a conception of justice that works as a regulative principle for a democratic regime. Defining justice, Rawls retains the idea of lexical priority, and the priority of right over the good. However, in defining goodness as rationality, Rawls is careful not to commit political liberalism to a comprehensive philosophical doctrine. Rawls states that it is incorrect to assume that no conception of the good can influence the political conception of justice, as the right and the good are complimentary (Rawls 2005, pp.173-174). In order to identify the size of the scope necessary to allow for various systems of belief and ways of life, justice must entail a perspective on the good; justice cannot be entirely neutral on all matters relating to the good. Certain conceptions of the good would be entirely at odds with the values of justice as fairness. However, the way in which conceptions of the good inform justice must be subject to restrictions. Rawls states:

a. That they are, or can be, shared by citizens regarded as free and equal; and
b. That they do not presuppose any particular fully (or partially) comprehensive doctrine (Rawls 1971, p.176).

Though justice is informed by the good, priority is always given to the right to ensure that the conceptions of the good people hold are compatible with justice as fairness. While we are free to develop our own conception of the good within the limits of what is just, there must be common features of the good that all reasonable people can agree on. Expanding on the thin theory of justice, there are five ideas of the good within justice as fairness, all of which are subsequent to the right:

1. The idea of goodness as rationality
2. The idea of primary goods
3. The idea of permissible comprehensive conceptions of the good
4. The idea of the political virtues
5. The idea of the good of a well-ordered (political) society (Rawls 1971, p.176).

Goodness as rationality is the idea that the citizens of a well-ordered society have rational plans of life, and that they arrange their actions around this plan (Rawls 2005, pp.176-178).

The primary goods, as noted in 2.1, grant individuals the ability to follow their rational plans of life. The idea of permissible conceptions of the good, and the idea of political virtue, refer to the range of ideas that inform these rational plans. Justice as fairness does not remain neutral between these ideas (Rawls 2005, p.192). Some ideas will be encouraged, while others will be discouraged if they fail to be compatible with the principles of justice (Rawls 2005, pp.195-200). Finally, Rawls states that a well-ordered society is good for two reasons: it allows persons to exercise their two moral powers and to secure their self-respect, and it enables a system of fair cooperation in which individuals can work together towards shared final ends (Rawls 2005, pp.202-204). These ideas delimit the conceptions of the good that individuals are able to formulate, and help identify the societal structure that will allow the citizens of a society to realise these goods.

As in *A Theory of Justice*, in *Political Liberalism*, "the basic structure of society is the first subject of justice" (Rawls 2005, p.257). The agreement of the social contract provides the principles that regulate it. Rawls posits that the role of the basic structure is to provide "background justice" while fair and voluntary social transactions are made within this setting (Rawls 2005, pp.265-267). Supporting the basic structure are the basic liberties, as indicated in the first principle of justice. Rawls views these liberties as necessary for allowing an individual to fully exercise their "two moral powers of personality over a complete life" (Rawls 2005, p.293). The two moral powers will be fully explained in the following chapter. These moral powers relate back to the idea of self-respect, and, as shown in 2.2 and 2.3, it is through self-respect that stability is maintained. Stability is integral to the idea of justice as fairness in Political Liberalism.

In Rawls' later work, justice as fairness has two characteristics: it is free-standing, and it is also self-sustaining (Rawls 1989, p.234; Rawls 2005, pp.207-212). It is free-standing as the idea of justice is derived only from political values: it is independent of religious, moral, and metaphysical values. It is self-sustaining as it generates its own support: citizens are motivated to accept the principles of justice through their self-respect. On finding that the principles of justice advance her ends, a person's sense of self-respect is enhanced. Thus, justice as fairness is stable. Once it is found to satisfy people's self-interest, people are inspired to maintain their commitment to it.

In sum, justice as fairness is shown to be a purely 'political' value in *Political Liberalism*. That is, justice is political rather than moral or metaphysical. A person can accept justice as fairness whatever their other beliefs as it is devised as not to conflict with other such beliefs. Providing a person is reasonable – that is, committed to beliefs that are compatible with the beliefs of other reasonable people – then that person can accept justice as fairness, seeking out her own ends within the scheme of permissibility as established through goodness as rationality. Once a person realises that she can achieve a successful plan of life in accord with her reasonable conception of the good, because she lives in a society that enables her to do so, justice as fairness becomes a self-sustaining doctrine. Reasonable people are motivated to continue supporting justice as fairness. It is these features of political liberalism that make it 'political'. Whereas a person committed to comprehensive liberalism also commits to certain other philosophical beliefs, political liberalism requires only a political commitment.

2.6 The Overlapping Consensus

This explains how Rawls defines justice as fairness in his later work. It remains to be seen how Rawls views individuals as coming to accept this conception of justice, and how it is a political conception of justice. Citizens determine political arrangements through the process of public reason, according to which they appeal only to reasons acceptable to other citizens. When citizens respect these arrangements, an overlapping consensus comes into existence. This consensus will now be examined.

As Rawls notes, "political power is always coercive power" (Rawls 2005, p.136). In a democratic regime, the justification for the use of political power rests on the will of the public. For the use of political power to be legitimate, it must receive continued public support. This raises the question of when the public would view the use of political power as justifiable. To this, Rawls responds that political power is justifiable when it is based on a constitution that is endorsed by free and equal citizens in light of their human reason (Rawls 2005, p.137). However, citizens in modern democratic societies do not necessarily share one, homogenous belief system, complicating the task of reaching a consensus allowing for the stability of a constitution. It is here that Rawls introduces the idea of the overlapping consensus.

The overlapping consensus is an arrangement whereby a range of reasonable comprehensive doctrines can cohere with a conception of justice (Rawls 2005, pp.144-145). It is through the agreement on a central conception of justice that they 'overlap'.

Through affirming the same conception of justice, citizens recognise that the overlapping consensus is in their own interest. Reasonable though opposed comprehensive doctrines can exist alongside one another. Part of a reasonable comprehensive doctrine must contain a recognition that other doctrines are reasonable. A person committed to a reasonable comprehensive doctrine does not attempt to use political power to suppress other reasonable comprehensive doctrines (Rawls 2005, p.61).

There are three features of the overlapping consensus: it is a moral conception, it is based on moral grounds, and it is stable (Rawls 2005, pp.147-148). Justice as fairness itself, the main object of the overlapping consensus, is a moral conception. Rawls does not explicitly state what is meant here, but in his explanation of the second feature it is implicit that the political conception of justice is also a moral conception of justice. That is, a certain sense of morality is entailed in the notion of justice as fairness. Though political liberalism avoids commitments to comprehensive doctrines, the overlapping consensus is in accord with the various comprehensive doctrines held by citizens. In this way, according to the second feature of the overlapping consensus, it is reached on moral grounds. Citizens come to endorse the political conception of justice through their commitment to their own comprehensive doctrines, including their sense of morality, as it is realised that their interests deriving from their comprehensive doctrines are advanced by the overlapping consensus. Furthermore, it is stable because it is not based on a temporary agreement between citizens. The citizens are committed to the political conception of justice alongside their other philosophical and religious commitments. They would not rescind their support for the political conception of justice, even in the event of their comprehensive doctrine becoming dominant within society. Seeing the opportunity to impose their doctrine on the rest of society, citizens would not seize it as they recognise that their interests are better served through the overlapping consensus.

Rawls' example of an overlapping consensus features three doctrines: a doctrine of free faith, a liberal moral doctrine in the mould of Kant or Mill, and a less systematically unified doctrine, one containing non-political values (Rawls 2005, p.145). Despite the tensions between these different modes of thought originating from the burdens of judgement, through their commitment to the principles of justice, Rawls shows how an overlapping consensus is achieved. As all three doctrines are reasonable, any one doctrine would not attempt to use state power to prohibit the other two. Each doctrine would also recognise the other two doctrines as being reasonable; as being compatible with a socially cooperative world. Through publicity, it would be recognised that the principles of justice enabled a mutually beneficial social environment that protected their rights and liberties, ensuring that they were able to advance their own ends in accordance with their comprehensive doctrines. All three doctrines are, then, containable within the overlapping consensus.

The overlapping consensus is built upon what Rawls terms the "constitutional consensus" (Rawls 2005 p.164). Where there is consensus on the basic liberties and democratic procedures, the constitutional consensus becomes an overlapping consensus, which has both depth and breadth. Depth relates to public discourse beyond that of the basics of the constitutional consensus, and of the citizens' own comprehensive doctrines (Rawls 2005, pp.165-166). It refers to how citizens come to justify their own political conceptions. Breadth relates to factors that ensure the freedom of all citizens to fulfil their political and social lives. Rawls argues that there is a need for legislation that guards freedom of thought and conscience, but also to protect material well-being, as when well-being falls below a certain level, it becomes impossible for citizens to be politically and socially active. Through the overlapping consensus, it becomes possible for a society containing a number of comprehensive doctrines to be well-ordered.

For Rawls, the well-ordered society consists of three elements (Rawls 2005, p.35). Firstly, it is centred on principles of justice that everyone knows and accepts, and that everyone is also aware that everyone else knows and accepts. Secondly, its basic structure complies with the principles of justice, and this is also known by citizens. Thirdly, citizens both endorse and comply with the principles of justice. With these three conditions satisfied, a well-ordered society can be established. Within a well-ordered society, unreasonable comprehensive doctrines that are incompatible with the overlapping consensus will eventually cease to exist (Rawls 2005, pp.195-201). Conceptions of the good that threaten to undermine the principles of justice are not compatible with political liberalism. Society is justified in using the coercive power of the state to discourage these doctrines. Following Isaiah Berlin, Rawls states that there is "no social world that does not exclude some ways of life" (Berlin 1990; Rawls 2005, p.197).

Political Liberalism further developed the conception of justice Rawls first explicated in *a Theory of Justice*. Retaining the central idea of justice as fairness, and the basic elements of the original theory of justice – the original position, the veil of ignorance, primary goods, the two principles of justice, lexical priority, and democratic equality – Rawls revised these ideas in order to respond to the realities of modern democratic societies. As these societies contain a variety of different belief systems, Rawls sought to identify a formulation of justice which could be endorsed by all reasonable persons, and that could contain all reasonable comprehensive doctrines. Rawls provides two key ideas in response: public reason and the overlapping consensus. Public reason provides the means by which citizens are able to advance their own ends. The overlapping consensus is a framework in which reasonable comprehensive doctrines exist together in a society. Though these doctrines may not be in accord with one another on specific religious, moral, or metaphysical grounds, they are in accord in their endorsement of justice as fairness. This is due to justice as fairness being a political conception of justice. As it is not dependent on wider philosophical commitments that would preclude other beliefs, reasonable persons can support justice as fairness without rejecting the beliefs to which they are committed. With these conditions satisfied, a well-ordered society that is also stable can function. Other aspects of political liberalism regard Rawls' conception of the person and autonomy, though these will be more fully explained in the subsequent chapters. For now, a basic outline of Rawls' argument in political liberalism has been provided.

2.7 Political Liberalism and Contemporary Political Thought

Kukathas states that Rawls' formulation of a liberal conception of justice has been recognised by both intellectual allies and opponents as the most substantial work of its kind for some time (Kukathas 2003, p.3). Given its influence, Rawls' work has been subject to a number of criticisms. Rawls' theory of justice has been critiqued from right-libertarian, communitarian, Marxist, and utilitarian perspectives. Robert Nozick and Michael Sandel have provided two particularly prominent critiques of Rawls' theory (Nozick 1974; Sandel 1984). However, I focus here on responses made to the doctrine of political liberalism. Specifically, I look at criticisms pertinent to the critique devised in this thesis. I begin by examining critiques of Rawls' ideas and then assess their strength. First, I focus on criticism of the idea of the comprehensive doctrine, problems with the overlapping consensus, and Rawls' underestimation of the problem of disagreement. Second, I examine the role of truth in political liberalism. Third, I assess G.A. Cohen's argument for the need of an "egalitarian ethos". Fourth, I examine Samuel Scheffler's argument that political liberalism has limited applicability. Finally, I assess the merits of these arguments, preparing for their development in relation to the critique devised in this thesis.

First, for Scheffler, political liberalism does not give an accurate account of the range of comprehensive doctrines contained in a modern democratic society. The example of an overlapping consensus referred to in 2.6 – a doctrine of free faith, a comprehensive liberal moral doctrine, and a partially comprehensive doctrine (Rawls 2005, p.145; Scheffler 1994, p.8) – does not represent the range of diverse beliefs in a modern democratic society. Though these three doctrines may cohere in an overlapping consensus, this does not imply that the multiplicity of doctrines in a modern society could be contained in such a consensus. In addition, Scheffler states that in endorsing the principles of justice through their own comprehensive doctrines, citizens will also be forced into maintaining incoherent beliefs. Utilitarians would endorse the principles of justice through their comprehensive doctrine of utilitarianism, yet simultaneously reject utilitarianism through the principles of justice (Scheffler 1994, p.9)

Iris Marion Young argues that Rawls' idea of the comprehensive doctrine is "too thin" (Young 2003, pp.183-185). Very few people live their lives strictly according to a comprehensive doctrine; philosophical systems tend to be too abstract to give meaning to everyday life. Instead, the values that influence people's beliefs and shape their everyday lives generally originate from a variety of sources. Young states that only a religious fundamentalist would live their lives according to a single comprehensive doctrine (Young 2003, pp.183-185). With people living their lives according to a number of different values and beliefs, Young argues that there may be overlap between these beliefs, but there is little chance for consensus. For Young, Rawls underestimates the depth of the disagreements inherent in modern societies.

Tim Hurley's criticism of political liberalism follows on from those of Scheffler and Young. Whereas Scheffler and Young critiqued the idea of the comprehensive doctrine, Hurley focuses on what he views as Rawls' inadequate solution to the problem of disagreement. Examining the burdens of judgement, Hurley argues that Rawls does not provide an adequate response to the problem of reasonable disagreement. Hurley states that while it is possible to disagree on questions of the good, Rawls would argue that any disagreement on the right was unreasonable (Hurley 2003, p.45). For the well-ordered society to function, there must be agreement on the priority of the right over the good, which is then reflected in public reason and the overlapping consensus. Anything incompatible with this arrangement is unreasonable. There are two problems Hurley finds with the dichotomy between the reasonable and the unreasonable. The first is that people will not be acquiescent in rejecting their comprehensive doctrines, where these doctrines are judged to be unreasonable (Hurley 2003, pp.46-47), as Andrew Murphy also argues (1998). Within a comprehensive doctrine there will be an account of what is and is not reasonable, and this may differ from Rawls' account.

Hurley's second criticism bring us to the second line of criticism assessed here: the relationship between the truth and political liberalism. As Hurley notes, Rawls is not concerned if the true comprehensive doctrine is rejected by other reasonable comprehensive doctrines. In formulating justice as fairness, the truth is not of importance for Rawls. Due to this absence of truth, Rawls would be left on shaky ground in the event of a comprehensive doctrine rejecting his account of the reasonable, as there is no recourse to the truth. The second of Hurley's problems relates to what he calls the "doctrine of conscientious refusal" (Hurley 2003, p.47). Hurley argues that while some societies have instances where they respect the right of individuals to follow their own conscience in contradiction to what the law demands of them – such as conscientious objection to military conscription – these cases are rare. However, in Rawls' theory, this idea plays a central role. Citizens will be able to disobey the law whenever their reasonable comprehensive doctrine gives them reason to do so (Hurley 2003, pp.47-48). Thus, even in cases where the law is reflective of the truth and human reason in general, if a citizen has reasonable grounds for disobeying it in light of their reasonable comprehensive doctrine, they will be justified in doing this.

Joseph Raz also argues that Rawls' position on truth in political liberalism exposes its weaknesses. Raz argues that political liberalism is reliant on what Raz terms "epistemic abstinence" (Raz 1990, p.4). Coming to a theory of justice without committing to metaphysical or moral doctrines means that justice will lack epistemological foundations. Without these foundations, a theory of justice cannot hold a position on the nature of truth. This would be to assert that justice could exist without truth, which as Raz argues, is evidently untrue (Raz 1990, p.15). To claim that something is just, there must be a standard of truth against which the claim can be measured. Yet this does not lead Raz to reject political liberalism. Instead, Raz argues that the overlapping consensus should be reconfigured, and that political liberalism requires a standard of truth. This truth is that society requires a fair system of cooperation (Raz 1990, pp.17-18). Therefore, the overlapping consensus, and the stability to which it leads, are necessary, as without them, a fair system of cooperation is impossible. If this is the standard against which we judge something to be just, and this standard can be turned from theory into practice, then we can say that this theory of justice is true. Justice, for Raz, cannot be independent of the truth.

Raz's line of criticism is similar to Jean Hampton's. While not rejecting Rawls' overall argument for political liberalism, Hampton argues that political philosophy need not be separated from metaphysics (Hampton 1989, p.813). Hampton distinguishes between "Socratic philosophising" and the sort of discourse Rawls has in mind when discussing public reason (Hampton 1989, p.808). Socratic philosophising is a method whereby matters of truth can be deduced through reasoning. According to Hampton, Rawls would allow for Socratic philosophising in the private realm but not in the political realm. Whereas in other schools of philosophy – ethics, aesthetics, and philosophy of science – matters of truth are of importance, in political philosophy, according to Rawls, truth is not the aim. When deciding on matters of justice through public reason, Socratic philosophising on the nature of truth must be avoided. For Hampton, this is an error. There are certain issues that have political significance, but that are impossible to discuss without invoking moral doctrines that incorporate a stance on truth. The legal status of pornography is such an issue according to Hampton (Hampton 1989, p.810). Some would argue that pornography is immoral; this might be due to religious reasons or a feminist stance against pornography. Others may argue that laws against pornography violate freedom of speech. If such an issue was to be approached through public reason, Hampton argues that allowing the expression of personal moral views would be beneficial. The expression of these views would allow people to revise or change their own opinions in light of the strength of the arguments. Hampton posits that in matters of tolerance, it is tolerance of people that is of importance, not tolerance of ideas (Hampton 1989, p.811). People may vehemently reject ideas, but there is nothing objectionable about this if these people are tolerant of those who hold these ideas. The nature of ideas expressed within public reason should not concern us, and the allowance of metaphysics within public reason would grant the conclusions reached greater depth.

Third, G.A. Cohen criticises the inequalities Rawls considers just on the difference principle. Part of Cohen's criticism hangs on Rawls' primary focus on the basic structure of society. Justice, for Cohen, requires more than this. For the kind of justice that Rawls envisions, Cohen argues that an ethos, or a "culture of justice", is necessary (Cohen 1992, p.315). People in such a society need to be motivated by this ethos. To illustrate this, Cohen describes two interpretations of the difference principle: one strict and one lax. On the lax reading, while the basic structure of society functions in accord with the principles of justice, people do not necessarily act from these principles. Thus, while I might respect the difference principle being applied at the societal level, in my everyday life, I may act contrary to the principle. If I am offered a job with a high salary that will exacerbate inequalities, I can justify this by saying that justice does not require me to devote my life to the less well-off, and I can still pursue my self-interest while acting justly (Cohen 1992, p.313). This lax reading of the difference principle, Cohen claims (1992, p.315), draws an arbitrary line between self-interest and the needs of others. Cohen argues that Rawls could not defend such an interpretation of the difference principle, as it is representative of an imperfect balance rather than a "fundamental principle of justice" (Cohen 1992, p.315). This leads instead to the strict interpretation. On this interpretation, more is required than governmental implementation. Rather, we are all required to cultivate a sense of justice, acting from this sense of justice in our everyday lives; what Cohen calls the "egalitarian ethos" (Cohen 1992, pp.315-316). The person acting from this ethos could justify her acts through reflection on the principles of justice, though Cohen stresses that constant appeal to these principles is unnecessary. Instead, the principles of justice are internalised and come to inform an individual's acts. For a society to be just, it must be characterised by less equality than a lax application of the difference principle would allow.

Fourth, as Rawls conceded, political liberalism cannot be universally applicable (Rawls 1993; Scheffler 1994, pp.20-22). Political liberalism is dependent on the existence of certain liberal traditions and institutions. There is no reason to suppose that societies without liberal values would adopt them. Thus, Scheffler argues that political liberalism has little to offer aspiring democracies. If political liberalism is then only applicable to a few modern liberal democracies, Scheffler posits that its defence of liberal principles is "intolerably weak" (Scheffler 1994, p.21). Because Rawls was writing in defence of liberal democracies, there is little to offer in response to this objection. If people are to be convinced by Rawls' arguments, they must already find something of value in existing liberal institutions; their moral psychologies must sit roughly in accord with what Rawls expects of citizens.

I now turn to assess the validity of these critiques, drawing out aspects that are further explicated throughout this thesis. First, the comprehensive doctrine and the range of disagreement in modern societies. In response to Scheffler's first argument, it could be argued that the range of comprehensive doctrines is of no consequence. What is of importance is that the comprehensive doctrines affirm the principles of justice. Even a society containing many different doctrines that stand in complete contradiction to one another could endorse justice as fairness, providing the doctrines are reasonable. Doctrines that are unreasonable and incompatible with justice as fairness would eventually cease to exist Rawls (2005, pp.195-201), as a democratic society that adopted justice as fairness would be justified in discouraging such doctrines. Nevertheless, Scheffler draws our attention to the level of disagreement in modern societies, which is also developed in Young's critique. A response to Scheffler's second argument is more problematic. That a utilitarian would accept justice as fairness through utilitarianism and then reject utilitarianism through the same principle seems an inevitable fact of political liberalism. This is, then, a powerful critique of political liberalism, especially regarding doctrines with political aspirations. While it is conceivable that religious groups may not wish to use state power to advance their ends, it is unlikely that those who hold deep political convictions would not desire to use the mechanisms of the state to further their aims.

A similar response to Scheffler's first argument could be made to Young's critique. While it is true that Rawls does not accurately reflect the beliefs people in modern societies hold, Rawls states that the comprehensive doctrine can be fully or partially comprehensive (Rawls 2005, p.152). The idea is not designed to anticipate every possible belief in a modern society. The content of the comprehensive doctrine matters little if the doctrine is compatible with justice as fairness. Despite this, Young and Scheffler both raise the problem of disagreement in modern societies. In a society deeply divided by many conflicting sets of belief, it is highly optimistic to imagine the citizenry could all hold one conception of justice. As mentioned in the Introduction, in addition to disagreeing with one another, many people in modern democracies hold little faith in the institutions that uphold liberal ideals. Furthermore, as Hurley notes, people may not willingly adjust their ideals should they be found to contradict the principles of justice. The task of identifying a conception of justice that everyone endorses is perhaps a more demanding one than Rawls assumes. This problem is returned to in later chapters.

Second, the role of truth in political liberalism. Jonathan Quong responds to Raz by stating that there is a difference between metaphysical and mundane truths (Quong 2011, pp.226-229). If the truth plays any role in political liberalism, it is only the mundane kind of truth. To state that if a theory of justice establishes what it set out to establish then it can be considered true is only a mundane truth claim. The important aspect of this for political liberalism is that the theory of justice does not preclude other philosophical and religious doctrines. Quong argues that Raz does not show this to be the case. If we consider justice as fairness as being true, this truth is mundane, not metaphysical. However, as I argue in Chapter 3, this overlooks distinctions between different types of truth claims that are not merely mundane. Such claims have a place within political liberalism. Rawls argues that scientific knowledge, where it is well-established and not controversial, can be drawn out to offer support to public reason (Rawls 2005, p.67). Thus, certain types of truths are included within political liberalism. How to determine what a scientific truth is as opposed to a metaphysical truth is a less straightforward task than Rawls appears to assume. This is further developed in Chapter 4; I draw on the arguments of Raz and Hampton to support the position that metaphysical claims are sometimes necessary within politics.

Third, the egalitarian ethos. In Chapter 9, I argue, in agreement with Cohen, that such an ethos is necessary for a just society, though for different reasons. Rawls is dependent on the truth of free will and a particular conception of moral psychology; we must be able to freely choose a conception of the good, while ensuring this conception remains within the bounds of justice. To do so, we must be considered as possessing a degree of control over our thoughts. If we do not have free will and do not exercise full control over our thoughts in this way, something else is needed to ensure the stability of a just society. This is why an ethos of justice is necessary.

Fourth, the applicability of political liberalism. I agree that political liberalism has limited applicability outside of a relatively small number of contemporary liberal democracies. Without pre-existing democratic institutions and a history of liberal ideals that have influenced the moral psychologies of citizens, there is little reason for someone to be persuaded to endorse political liberalism. However, this is not only a problem for post-authoritarian regimes. As both Scheffler and Young note, Rawls underestimates the level of disagreement in modern democracies. In addition, I argue that Rawls assumes a degree of commitment to democratic institutions on the part of citizens that is not necessarily held. As mentioned in the Introduction, many citizens in modern liberal democracies may be persuaded of the superiority of authoritarianism. Leaving citizens to determine their own good may lead them to reject the value of liberal ideals. This is further reason for the need of something akin to the egalitarian ethos. In later chapters, I argue that Rawls' formulation of full autonomy provides a counterpart to this ethos.

While political liberalism has been critiqued in several ways, these four points are pertinent to the critique developed here. First, political liberalism is highly optimistic. As Young notes, Rawls underestimates the extent to which modern societies are divided, and the types of disagreement they are divided by. Second, the relationship between the truth and political liberalism. This is more complex than Rawls assumes. Certain truth claims must be accepted within any political regime. Fourth, political liberalism has limited applicability. Societies without democratic traditions and liberal institutions are offered little to persuade them of why it is preferable. This problem, however, stretches beyond authoritarian regimes. If citizens in modern democracies fail to be convinced by liberal ideals, political liberalism does not offer a solution.

2.8 Conclusion

Throughout his work, Rawls' aim is to defend his theory of justice: justice as fairness. In *A Theory of Justice*, defines this theory, intending it as a regulative principle for modern constitutional democracies. Modern societies, however, are divided by many different sets of belief systems. Critics of *A Theory of Justice* posited that this theory was incompatible with this plurality of beliefs. In *Political Liberalism*, Rawls attempted to reformulate justice as fairness to make it compatible with the range of comprehensive doctrines contained in modern societies. Concepts such as goodness as rationality were reformulated, while public reason and the overlapping consensus were devised to show how people within modern democracies could unite around a common understanding of justice without rejecting their own beliefs. Through these concepts, justice as fairness is conceptualised as a free-standing theory of justice, which is not committed to philosophical, religious, or moral concepts, and is compatible with all reasonable comprehensive doctrines.

I have identified four criticisms that of political liberalism that are developed throughout this thesis: the range and types of disagreement in modern societies being underestimated; the relationship between the truth and politics being more complicated than imagined; the need for a stronger commitment on the part of citizens, and to importance of justice in everyday life; the need for prior commitment to the value of democratic institutions and liberal ideals.

The purpose of this chapter has been to explain the basic argument running throughout Rawls' thought, with particular attention given to *Political Liberalism*. In subsequent chapters, I define Rawls' political conception of the person and the concepts of rational and full autonomy. With an understanding of the basic argument of political liberalism, the relationship between these ideas and the overarching aim of Rawls' work will be understood more clearly.

Chapter 3 – Autonomy and the Person

3.1 Introduction

The "political conception of the person" is Rawls' attempt to explain the basic features of human nature and psychology without committing to a comprehensive view of human nature (Rawls 2005, p.29). Rawls' theory of autonomy is supported by this conception of the person. This chapter explains the conception of the person and Rawls' theory of autonomy. For the state and society to function according to the ideals of political liberalism, at least some citizens must be fully autonomous. Attainment of autonomy is dependent on the development of first moral powers as to recognise justice as fairness. I argue that the state requires the moral development of its citizens to accord with the Rawlsian values of fairness and reasonableness, but in order to achieve this, the state's education system becomes partially perfectionist. This argument is further developed in Chapter 9. Though Rawls attempts to avoid commitments to comprehensive doctrines, these commitments are unavoidable in regard to moral education. Furthermore, within these commitments is, I argue, an unacknowledged commitment to the truth of free will. I develop this argument across this chapter and Chapter 5. To begin, it will be necessary to understand how Rawls thinks about human nature and psychology in *Political Liberalism*, how he tries to avoid commitments to metaphysical doctrines, and the level of his success in this task. In Chapter 5, I then show how within this way of thinking about the person is a commitment to the truth of free will.

Other theories of moral psychology, as found in the work of Kohlberg and Haidt, are assessed in this chapter. Through this assessment, I draw out the importance of the group and the environment in shaping our autonomy and moral psychology. While Rawls thought political liberalism was compatible with any reasonable theory of human psychology, I argue that Kantian moral psychology undergirds Rawls' thought.

In this chapter, I define the political conception of the person in 3.2, discussing the role it plays within political liberalism. Both rational and full autonomy are explained in 3.3. I explore the dilemma of moral education in political liberalism in 3.4. Following M.V. Costa and Eamonn Callan, I argue that in order to develop autonomy in citizens, moral education must necessarily become partially perfectionist. The dominant theories of moral development are explained in 3.5. In 3.6, I further develop the argument for why aspects of perfectionism are necessary, using these theories for support. Throughout this chapter, I explain why the fully autonomous citizen is necessary for political liberalism, assessing some problems arising from psychology for the realisation of full autonomy.

3.2 The Political Conception of the Person

Rather than a political philosophy stemming from an understanding of human nature, as in the case of Hobbes (Macpherson 1985), political liberalism is a free-standing construct that is designed to work independently of psychological theories. Basic features of human nature and psychology place limits on what can be proposed in political philosophy, but they do not determine the conclusions (Rawls 2005, p.87). All that is required of conclusions in political philosophy is that they do not contradict certain well-known basic aspects of human nature. In the political conception of the person, Rawls states what underlies a person's moral psychology, and their basic motivations. Claims that go beyond this limited description, such as what individuals' main objectives in life should be, or how individuals should act or generally behave, are avoided. For instance, the political conception of the person is not an Aristotelian perspective on man as a political animal, as such a view presupposes a comprehensive view of human nature, one that prescribes how humans should act across their lives. As Soran Reader notes, Aristotle's conception of the person includes a metaphysical perspective on human nature (Reader 2007, p.581). Instead, Rawls is only interested in the basic features of human nature as would generally be accepted, attempting to reveal their compatibility with justice as fairness. In this section, I explain these basic features, assessing how Rawls constructs the political conception of the person.

This understanding of human nature differs from that which was offered in *A Theory of Justice*. Thomas Baldwin argues that the account of human nature in *A Theory of Justice* was premised on a comprehensive understanding of psychology (Baldwin 2008, pp.248-249), in which certain characteristics of human nature are considered essential for a fulfilling life. These characteristics include the necessity of behaving in accordance with principles of justice, and feelings of guilt and shame resulting from a person acting against these principles (Rawls 1971, pp.440-446). As it incorporates a view on how humans generally *do* and *should* behave, this effectively constitutes a comprehensive moral doctrine, the type of moral philosophy Rawls wishes to avoid committing to in *Political*

Liberalism. Thus, in *Political Liberalism* Rawls refined this conception of human psychology.

At the centre of this refined understanding of human nature is the claim that all individuals have two moral powers: the ability to form a sense of justice and a conception of the good (Rawls, 2005, p.19). This, in combination with the human capacity for reason, enables people to be free. People desire to be free due to their possession of the second moral power. They wish to be free to live a life of their own choosing. Because they have these powers, people are capable of living in a cooperative well-ordered society. Their capacity for a sense of justice allows them to ensure that the terms of cooperation in a society are fair, and their ability to form a conception of the good enables them to pursue their own aims in life, formulating such aims in accord with the fair terms of cooperation. The person's social existence is of great importance for Rawls. Reader notes that Rawls' basic definition of the person is "someone who can take part in social life" (Rawls 2005, p.233; Reader 2007, p.585). While Rawls' understanding of the person is different from Aristotle's, the social world is still essential for Rawls. On the political conception of the person, a person must be capable of acting cooperatively in a society. This basic understanding of human nature informs Rawls' idea of the political conception of the person.

Through the political conception of the person citizens think of themselves as being free. It is this understanding of their own freedom that motivates citizens to live in a free, democratic society. According to Rawls, they think of themselves as free in three regards. First, they do so because of the moral powers, particularly the ability to form a conception of the good (Rawls 2005, p.30). They will pursue this conception across the course of their life through having a "rational plan of life" (Rawls 2005, p.177). Once a person has formulated a conception of the good, they may also continually revise how they conceive of the good. Changes made have no bearing on the person's political or institutional identity. People are free to decide for themselves what constitutes the good life, living accordingly. This is the first aspect of a person's freedom.

The second way in which people think of themselves as being free relates to social and political institutions (Rawls 2005, pp.32-33). People think of themselves as being the authors of valid claims. These claims are self-authenticating: they are valid because a rational and reasonable person made them. Because people can make valid claims, they view themselves as being able to make claims on social and political institutions in order to advance their own conceptions of the good. Rawls contrasts this aspect of freedom with the lack of freedom of the slave. A slave is not able to make claims on their institutions as a slave has no political rights. Slaves are therefore not viewed as being the authors of valid claims. For all citizens to be free in a liberal democracy, the state must ensure that all persons are able to make claims on their institutions. Citizens can, then, pursue their aims politically, ensuring that society is structured as to allow them to live lives of their own choosing.

The final aspect of freedom relates to a person's ability to take responsibility for their conception of the good (Rawls 2005, pp.33-35). As rational and reasonable agents, citizens ensure that their conception of the good is compatible with the political conception of justice; such goods cohere with a society based on a system of fair cooperation. They are therefore capable of realising that, when pursuing their conception of the good, they must be able to provide for what they expect to receive in return. When making claims on their institutions, they will also realise that their claims will be assessed in relation to their reasonableness. Citizens take responsibility for their conception of the good by ensuring it is a reasonable conception. Such a conception is reasonable because it does not challenge other citizens' freedom to formulate a similarly reasonably conception of the good.

As they think of themselves as being free in these regards, individuals desire to live in a society that allows for their freedom. Rawls' understanding of the basic features of human nature leads him to state what people would think of as being just within a society, and what would be sought to achieve this sense of justice. This relates back to the idea of the primary goods discussed in Chapter 2. As there are certain innate characteristics of individuals, there are certain goods that they will need to satisfy needs deriving from these characteristics. Rights and liberties will be necessary to allow people to obtain these goods.

Self-respect is also part of the person's basic psychology, and constitutes the most important primary good (Rawls 2005, pp.81-82). It is through self-respect that the stability of society as a system of fair cooperation is achieved. Self-respect provides the self-generating support of a society. Though little is said about how self-respect relates to the political conception of the person, Rawls later says a person who had no interest in developing their conception of the good, not caring about her basic liberties, would show a lack of self-respect. Such a person not be considered a "full person" (Rawls 2005, pp.76-77). While less is said in *Political Liberalism* about the types of virtues society should aim to cultivate, Rawls nevertheless maintains the belief that people should wish to live as part of a cooperative society, developing the necessary skills to achieve this. A person with no such desire has some kind of moral defect.

To summarise, the two moral powers are at the centre of Rawls' thought on human nature. They also provide the basis for the political conception of the person, through which people think of themselves as being free in three regards: free to form a conception of the good, living a life in accord; free to make claims on their institutions in light of this conception; free to take responsibility for this conception. Other ideas regarding human nature are carried over from A *Theory of Justice*, such as the idea of self-respect. Throughout all of Rawls' thought on human nature in *Political Liberalism*, an attempt is made to avoid making commitments to comprehensive doctrines. The political conception of the person is not designed to be dependent on broader psychological or philosophical theories of human nature.

3.3 Autonomy: Rational and Full

There are two forms of autonomy distinguished in political liberalism: rational and full. Rawls' thought on these forms of autonomy predates *Political Liberalism*, as he devised them in the paper 'Rational and Full Autonomy' (Rawls 1980). Rational autonomy relates to the citizen's self-interest and ability to for a conception of the good (Rawls 2005, pp.72-73). Parties in the original position are considered rationally autonomous; it is the citizen in the well-ordered society who becomes fully autonomous. Through their rational autonomy, parties in the original position will have established principles that allow for citizens to realise their self-interest. The citizen who acts *from* these principles is considered by Rawls to be fully autonomous (Rawls 2005, pp.77-78). Rawls' thought on autonomy follows the same logic as his thought on the conception of the person. The theory of autonomy must not be conditional on the truth of a comprehensive philosophical doctrine. In contrast with the comprehensive liberal doctrines of Kant and Mill, Rawls set out to establish a conception of autonomy as a purely political value. This constraint informs both rational and full autonomy. I begin by defining rational autonomy before moving on to full autonomy. The existence of fully autonomous citizens is, I argue, a prerequisite of political liberalism.

Rational autonomy is based upon a person's "moral and intellectual powers" (Rawls 2005, p.72). To help define rational autonomy, it will be necessary to understand what Rawls means by 'rational'. As discussed in Chapter 2, for Rawls, rationality refers to self-interest. It is rational to want certain goods and rights, as it is in a person's self-interest to want these things. Rational autonomy relates to what would be sought by parties in the original position in order to satisfy self-interest. A person is then deemed to be rationally autonomous if she possesses the ability to use her moral and intellectual powers to further her rational self-interest.

In addition to the two moral powers, there are two further conditions necessary for rational autonomy. First, parties must be able to decide upon fair terms of cooperation (Rawls 2005, pp.72-73). There is no external standard by which to judge the fair terms of cooperation; people must have the capacity to decide what is fair for themselves. Their ability to do this allows for them to be considered rationally autonomous. Second, because individuals have two moral powers, they also have two higher-order interests in developing these powers (Rawls, 2005, pp.73-74). Rawls argues that a person must have these moral powers, and be capable of developing them, in order to function as a cooperative member of a well-ordered society. This relates back to Rawls' views on self-respect; people who are not concerned with developing their moral powers would show a lack of self-respect (Rawls, 2005, pp.76-77), as they have no values upon which to base a rational plan of life, and no sense of morality to guide them. If the parties in the original position did not account for these higher-order interests, they could not be considered rationally autonomous. There is also a third higher-order interest: the parties must want to secure those they represent the conditions necessary for realising their moral powers. Due to the second moral power, people have the ability to form a conception of the good. Parties in the original position will therefore be motivated to ensure that the principles decided on allow for citizens to develop this conception, and to live according to it (Rawls 2005, p.74).

These three higher-order interests, along with the capacity to determine fair terms of cooperation, allow for rational autonomy. Rawls stresses that rational autonomy, as opposed to full autonomy, is applicable to the original position (Rawls 2005, p.75). In the original position, it refers to the rationality of parties, who are "artificial" representatives of citizens. Parties are rationally autonomous as they understand that citizens have certain self-interests deriving from their second moral power, and, in their deliberations, parties are motivated to secure these interests. Citizens are rationally autonomous when they come to exercise their second moral power. He also asserts that rational autonomy is not a matter of *pure* self-interest (Rawls 2005, pp.76-77). Though citizens are assumed to have self-interests, they are not *only* self-interested. They must also recognise others as having their own self-interests. Their self-interests might also be of an altruistic nature. This is not a matter of egoism. Parties are not aiming to maximise gains for those they represent with no regard for others. The aim of rational autonomy is also not to secure material goods, but to obtain the necessary conditions for citizens to develop their moral powers. It may be that the citizens a party represents are neither egoistic nor driven by the need for materialistic goods. In turn, with these conditions obtained, a citizen enhances her self-respect when she is able to develop her moral powers. This leads to the self-sustaining stability of a well-ordered society. As Rawls states, in order to obtain these conditions, principles of justice decided on in the original position must be conducive to this aim.

Full autonomy is also dependent on aspects of the original position, but it is not parties in the original position who are considered fully autonomous, it is the citizens of a well-ordered society (Rawls 2005, p.77). Citizens become fully autonomous when they act in accordance with the principles of justice determined in the original position.[1] Rawls argues that full autonomy requires not only compliance with the principles, it requires that citizens act *from* these principles. As citizens recognise the principles as being just, they are motivated to apply the principles in their political lives. This relates back to the idea of publicity. Through publicity, citizens recognise that justice as fairness enhances their self-interest, and are thus motivated to act according to the principles of justice. Therefore, full publicity must be attained in order to satisfy the conditions of full autonomy (Rawls 2005, p.78). Once a citizen understands and acts in accord with the principles of justice, she can be considered fully autonomous. With full autonomy realised, citizens enjoy the rights and liberties of a well-ordered society; they can participate in that society's collective self-determination (Rawls 2005, pp.77-78). The fully autonomous citizen is able to develop the two powers, living a rational plan of life according to her conception of the good, revising this conception when she desires to.

[1] Hence why parties in the original position cannot be considered fully autonomous. As the principles are not yet determined, parties cannot act from them.

Full autonomy is essential to political liberalism. If citizens did not act as fully autonomous citizens should, political liberalism would be left highly unstable. For the values of political liberalism to be realised, there needs to be people who are motivated to realise them. Thus, it is necessary that at least some citizens act as fully autonomous citizens.

While rational autonomy can be thought of as comparable to the ways in which Joseph Raz (1986) and John Stuart Mill (1859) defined autonomy, full autonomy is closer to the Kantian formulation. For Raz and Mill, autonomy is a matter of self-realisation. As Raz puts it, "the autonomous person is a (part) author of his own life" (Raz 1986, p.369). Similarly, a person is rationally autonomous if she is capable of determining the good for herself, living a life, then, of her own choosing. Rational autonomy is, however, a thinner understanding of autonomy than that identified by Raz. On Rawls' definition of rational autonomy, a person could choose to live a life of voluntary servitude and still be thought of as rationally autonomous. She has made a rational decision in light of her conception of the good. Her autonomy remains, then, intact. For Raz, a person needs an adequate range of options to be considered living an autonomous life (Raz 1986, p.373). Full autonomy, on the other hand, is closer to Kant's notion of moral autonomy. The morally autonomous person, for Kant, is the person who acts from universally applicable moral laws. There are two important aspects here. First, the autonomous person, for Kant, acts from principles that would be chosen by the exercise of reason in the noumenal realm, as for Rawls, the fully autonomous citizen acts from the principles of justice that would be chosen within the original position. Second, such a citizen recognises that her own legitimate sphere of action ends where another person's begins. She will not, then, transgress the bounds of another person's autonomy. Likewise, the fully autonomous person acts from the principles of justice. In doing so, she acts reasonably, not attempting to impose her own doctrine on others.

The two forms of autonomy are, thus, incorporated within the scheme of political liberalism. There is, however, a tension between the two. Though she should develop her rational autonomy within the bounds of reasonableness – she should not wish to impose her conception of the good on others – a person could exercise her capacity for rational autonomy to live a life in which she never acted from the principles of justice. For example, she may decide that a life free of political concerns was her route to happiness. This person is rationally autonomous, but we cannot consider her fully autonomous. On the other hand, a person who had been conditioned to act in accord with the principles of justice from a young age, never having been offered the opportunity to act otherwise, would not meet the conditions necessary to be considered autonomous by Raz. She does not have a range of adequate options from which to choose, as she has no motivation to choose otherwise. Thus, while she is acting as a fully autonomous citizen should, Raz would not consider such a person autonomous. The tension between these two conceptions of autonomy is further explored later in this chapter.

In sum, through rational autonomy, parties in the original position ensure that fair terms of cooperation are decided on, so that citizens can develop their moral powers. This ensures that fair terms of cooperation are enshrined in the principles of justice. Through full autonomy, citizens come to act *from* these principles. A citizen who is motivated to act from these principles can be considered fully autonomous. At the root of Rawls' thought on both autonomy and the person is the claim that we have two moral powers, in addition to the capacity to develop these powers, which is realised through full autonomy. A tension exists between rational and full autonomy, however. A person's capacity for one may preclude the development of the other. Without fully autonomous citizens, the values of political liberalism cannot be realised. The tension between the two forms of autonomy is, then, highly problematic.

3.4 Moral Education in Political Liberalism

Later in this chapter, I explain how our moral development is influenced by the communities we belong to, and the environment in which we are raised. I explain this in order to develop the argument that autonomy, as Rawls formulates it, is subject to influences beyond the individual, which Rawls does not recognise. It should not be assumed that an individual can become fully autonomous if left to her own devices.[2] To begin, however, I examine how moral development could occur within the well-ordered society of political liberalism. While Rawls believed political philosophy need not be dependent on a comprehensive view of human nature, a particular sort of person inhabits the well-ordered society. This person has the ability to realise a sense of justice and a conception of the good; her self-respect is dependent on this ability's realisation. Without

[2] As Rawls admits when he notes that full autonomy can only be attained when the full publicity condition is reached (Rawls 2005, p.78). We must know what the principles are if we are to act justly. However, though Rawls recognises that external factors can influence the development of full autonomy, he does not offer a complete explanation of how this development is to occur.

the existence of the person possessing these attributes, political liberalism would face an existential threat. It must be ensured that this person does come into existence, that the person is able to develop her moral powers, and that this person is continually replicated across society. While in *A Theory of Justice*, Rawls explains how the process of moral development should occur, little is explicitly said about moral development in *Political Liberalism*. Nevertheless, *Political Liberalism* provides several mechanisms through which moral development could occur, which I assess in this section. In 3.5, I then explain the role of the group in shaping an individual's moral development, in order to draw out the complexity of moral development in the politically liberal state. To begin, I examine how Rawls explained moral development in *A Theory of Justice*.

As James Scott Johnston notes, in *A Theory of Justice*, Rawls thought that the purpose of education was to guide the moral development of an individual in accordance with the principles of justice, allowing the individual to realise her self-respect (Scott Johnston 2005, pp.205-207). Though he commits to neither, at least ostensibly, Rawls identifies two courses of moral education in *A Theory of Justice*: one based on empiricism and the other on rationalism (Rawls 1971, pp.458-462). The empiricist approach uses what Rawls calls "missing motives". To guide the individual towards the desired sense of morality, the empiricist supplies a motive to do so where that motive is missing. Punishment or reward may be utilised early in the process. Later on, the aim is to develop psychological inclinations to follow certain courses of action. Rawls associates this approach with utilitarianism ("from Hume to Sidgwick") and Freud. The rationalist approach sees no need to supply missing motives. Instead, there are innate emotional and intellectual capacities within the individual. The purpose of moral education is, then, to allow the individual the freedom to realise these innate capacities. This approach Rawls associates with Rousseau, Kant, Mill, and Piaget.

Though Rawls states that his ideas on moral education are compatible with both approaches, in what follows, his ideas appear to remain faithful to the rationalist approach. Three stages of moral development are identified. In the first stage, a person learns to respect the morality of authority through the love she has for her parents (Rawls 1971, pp.462-467). The morality of association is learned in the second stage (Rawls 1971, pp.467-472). During this stage, a person learns the moral rules that guide the behaviour of people within groups. The importance of fairness and cooperation are learned; people who follow the rules are more successful in achieving their aims. Finally, the morality of principles is learned (Rawls 1971, pp.472-479). In this stage, the individual comes to understand abstract moral principles. It is in this way that Rawls imagines the individual comes to understand the principles of justice, and to be motivated to act from them. While Rawls argues this scheme of moral education is compatible with both the empiricist and rationalist approaches, Rawls does not mention the need to supply missing motives. Throughout his description of this scheme, Rawls appears to imagine the process occurring naturally, with the educational scheme only allowing the individual to develop her innate capacities. Rawls notes how his approach coheres with that of Kohlberg (Rawls 1971, pp.461-462). The morality of association is similar to the conventional level of moral development, in which the individual comes to learn the importance of following the behaviour of the group. This is further explored in 3.5 and 3.6.

There is little said regarding education in *Political Liberalism*, as Johnston notes (Johnston 2005, p.204), though educational theorists such as Kenneth Strike have looked at how Rawls' thought could be used to support a theory of education (Strike 1994). Johnston charts how Rawls' thought on education developed from *A Theory of Justice* to *Political Liberalism* (Johnston 2005, pp.205-207). Education in *Political Liberalism* turns away from the moral towards the political. That is, the focus on a scheme of moral development is omitted. The ordering of social and political values takes precedence over the realisation of individual autonomy and other moral values. Instead of the education system's primary purpose being the realisation of future citizens' autonomy and self-respect, its purpose now is found in publicity. Through education, the child must become aware of the importance of being a citizen, and the rights and liberties that come with citizenship (Johnston 2005, p.209). Self-respect is not abandoned; it remains imperative, but Johnston argues that its priority over other goods is moderated.

Contra Johnston, I argue the realisation of full autonomy is integral to political liberalism. Without the existence of the fully autonomous citizen, political liberalism loses its stability. As there would be no one who both recognised, and acted from, the principles of justice, political liberalism would lose its support. This raises the question of how the fully autonomous citizen is to come into being without making autonomy's realisation the ultimate aim of education. Rawls states that the basic structure of society has a formative psychological effect on citizens (Rawls 2005, pp.269-271), although through the concept of public reason, the institutions that make up the basic structure should not promote comprehensive moral doctrines. Based upon the principles of political liberalism, a school would be able to teach moral principles in accordance with the political conception of justice, but would be unable to instruct pupils in metaphysical, moral, or religious doctrines, where such instruction was aimed at encouraging the child to endorse a particular view. Education would be confined to what is right rather than what is good. Though a child would be educated in the idea of justice as fairness, and the Rawlsian value of reasonableness, her moral development outside of the school would be dependent on the family and experiences in the private sphere. Values taught in the private sphere could conflict with those taught in the school. If the motivation to endorse values taught in the private sphere outweighed the motivation to support justice as fairness, political liberalism's continued support would be under threat. For children to develop into the functioning citizens of a well-ordered society that Rawls hopes for them to be, it would be necessary to recognise the influence of institutions such as the school in enabling moral development.

Instead, as M. Victoria Costa notes, Rawls leaves instruction in moral values to political and judicial institutions (Costa 2004, pp.5-6). Rawls states that the political conception of justice has to act as an "educator", but that this is achieved through the "public culture and its historical traditions of interpretation" (Rawls 2005, pp.85-86). Costa argues that instructing people in the idea of justice as fairness would be the equivalent of teaching a comprehensive doctrine (Costa 2004, p.7). On this approach, a child would be educated in the importance of fairness so that, as an adult citizen, she comes to accept justice as fairness. If a person were to be educated through the idea of justice as fairness, that person should come to view its principles as being morally right. Being motivated to act in accord with what is right, the child would come to see this as a good. Moral principles drawn from other philosophical doctrines that, though reasonable, exist in tension with justice as fairness – libertarianism or utilitarianism, for example – would be recognised as morally inferior to Rawls' principles of justice. There would be no reason for a person to endorse other philosophical doctrines. The diversity of modern societies, which Rawls sought to maintain, would be lost.

The Rawlsian response, as anticipated by Costa, is to point out the difference in scope between the comprehensive doctrine and the idea of justice as fairness. Costa argues that this is inadequate. Political liberalism's education system would be required to teach political and social values that conformed to the idea of reasonableness. These teachings would also have a deep effect on a person's character and moral development. The effect of this would be to mould an individual into being a liberal subject, which appears to bear little difference to the prescribed autonomy of Kant. This could preclude a person from desiring to endorse doctrines deriving from other philosophical or religious traditions, as even if the doctrine was reasonable, there may be little reason for a person with a fundamentally liberal moral psychology to wish to adhere to it.

Instructing children in the moral value of fairness, these institutions cease to be independent of comprehensive doctrines. The liberal ideal of fairness would come to hold great power over the child's development, prohibiting the child from desiring to accept alternative doctrines. Political liberalism then faces a dilemma. One option is to teach the values of justice as fairness in order that the moral powers are developed in accordance with this concept. With this approach we risk losing the diversity of moral doctrines within a society. The other option is to leave instruction in morals entirely to the private sphere, through which the moral powers may be developed in accordance with alternative belief systems. On this approach, there would be no reason to assume that a person whose moral powers had developed under an alternative religious or philosophical comprehensive doctrine could be reconciled to the idea of justice as fairness. Fairness may hold no value within this doctrine.

Due to Rawls' belief that the political conception of justice should act as an educator, it is fair to assume the former option should be chosen. However, as Eamonn Callan argues, education must be at least partially comprehensive, because of the strong effect it has on a person's development (Callan 1996, p.6). The potential result of this, Callan argues, could be that children reject the doctrines held by their parents, as they are unmotivated by ideals and values drawn from sources other than liberalism (Callan 1996, pp.22-23), a problem which Rawls recognised and saw as regrettable (Rawls 2005, p.200). As the purpose of political liberalism is to devise a conception of justice that is acceptable to all the reasonable systems of belief in modern societies, an education system that reduced support for these beliefs would undermine political liberalism's objective. For a citizen to become fully autonomous, however, this would be unavoidable. The fact of whether citizens are able to make reasonable claims on their institutions and take responsibility for these claims is dependent on the development of their moral powers. A person who had developed an entirely illiberal, unreasonable sense of justice and conception of the good would not be deemed capable of making reasonable claims on institutions, nor of taking responsibility for them. This person would therefore not be considered fully autonomous.

3.5 Theories of Moral Development

At the heart of Rawls' thought on human nature is the claim that we have two moral powers. Amartya Sen, agreeing with Rawls, argues that the presumption of the moral powers is central to "the tradition of democratic thought" (Rawls 2005 pp.18-19; Sen 1999, p.272); to deny this would be to depart from this tradition, and to undermine human rationality. In this section, I explore how moral development is understood in psychology in order to assess how we come to develop a sense of justice and a conception of the good. The purpose is not to endorse any one theory of moral psychology, but to draw out the importance of the group and emotion in determining the individual's moral development. This importance of the group is recognised in both the dominant strands of the study of moral development: in Lawrence Kohlberg's conventional level of moral development and in Jonathan Haidt's idea of the "moral community". I turn first to Jean Piaget and Kohlberg, then assess the work of Haidt. Following this, I assess Patricia Churchland's critique of Haidt. Though there is much disagreement between the various approaches, I argue that there is agreement on two matters: the importance of community and emotion in determining our sense of morality.

Study of moral development in psychology is commonly thought to begin with Piaget (Haidt 2008, p.65). Piaget argued that moral development occurred across stages. This theory was revised by Kohlberg, who posited six stages of moral development (Kohlberg and Hersh 1977, pp.54-55). The six stages come under three levels. During the preconventional level, the child considers only what is culturally considered to be good or bad, and their behaviour is guided by the avoidance of punishment, or the obtainment of hedonistic pleasure. At the conventional level, the family, group, and nation are considered important for their own sake; behaviour conforms to the expectation of the collective. Finally, at the postconventional level, individuals determine what is right according to abstract moral principles, which exist apart from the group of which the individual is a member. With the realisation of the postconventional level, the individual can deduce what is morally right through a process of moral reasoning. The following table explains each stage of development:

Preconventional Level	
Stage 1 – Punishment and Obedience Orientation	Avoidance of punishment provides the sole determination of whether an action is deemed to be good or bad. There is no respect for a moral order outside of the person's obedience to authority.
Stage 2 – The Instrumental-Relativist Orientation	Actions are deemed to be right if they satisfy one's own needs, or sometimes the needs of others. The satisfaction of needs is considered in pragmatic ways, occasionally in terms of fairness, but only insofar as fairness supports the obtainment of satisfaction.
Conventional Level	
Stage 3 – The Interpersonal Concordance Orientation	Actions begin to be judged by intention. It is important that one's actions are considered to be good by others, and behaviour should conform to the norms of the social order.
Stage 4 – The "Law and Order" Orientation	Actions should uphold the social order, and persons come to hold respect for authority and doing one's duty.
Postconventional, Autonomous, or Principled Level	
Stage 5 – The Social-Contract, Legalistic Orientation	Actions are considered in terms of what has been determined to be right by society as a whole through a process of critical examination. Consideration is given to the relativism of personal values, and it is thought that procedural rules should be followed in order to reach consensus.
Stage 6 – The Universal-Ethical-Principle Orientation.	Abstract universal principles guide moral behaviour. Justice, fairness, and equality are considered important for their own sake.

(Kohlberg and Hersh, 1977, pp.54-55)

Kohlberg thought that these stages of moral development were universal. To prove this, Kohlberg undertook cross-cultural empirical studies, interviewing participants across a number of years (Colby et al 1983; Gibbs 2013, p.86; Gibbs et al 2007). Unfortunately for Kohlberg's theory, the empirical studies revealed its inconsistencies. During college years, in which participants would be expected to be at the conventional level, 20% of Kohlberg's sample regressed to stage two. There were also problems with stages five and six (Gibbs 2013, pp.89-90). Stage five was rarely reached, and stage six was never reached. John Snarey argued that stage five was based on the philosophy of Kant and Rawls (Snarey 1985, p.228), being based as it is on procedural rules toward determining a consensus, and that there were difficulties with its cross-cultural translatability. Not all, or even many, cultures would ascribe to a Kantian notion of morality. Furthermore, Kohlberg admitted that stage six was based on an elite sample, which would only be recognised in persons with philosophical training (Gibbs 2013, p.90). Rather than being a universal process, in which individuals progress linearly from stages one to six, there is significant variation in moral development. Kohlberg underestimated the importance of the conventional level and the power the group has over the determination of its members' sense of morality.

Despite the inconsistencies in the stages of moral development, the theory of cognitive development has remained one of the primary understandings of moral development in psychology (Haidt 2008, p.68). An alternative understanding is found in Haidt's "new synthesis", which brings together evidence from social psychology, neuroscience and evolutionary theory (Haidt 2008.). Instead of attempting to understand moral development from the perspective of universal moral principles, the new synthesis looks to communities. The 'moral community' is the ultimate arbiter of morality, according to the new synthesis. E.O. Wilson, who originally invented the phrase "the new synthesis", thought that morality would become better understood through sociobiology (Wilson 2000), as human behaviour would come to be assessed through knowledge of neural and evolutionary processes. According to Wilson's theory, our brains have been shaped through evolution to experience moral emotions. It is these emotions that determine our intuitions regarding what is right and wrong.

Haidt used this biological evidence to posit a theory of moral development on which these emotional reactions are contingent (Haidt 2007). Verbal reasoning – the kind of reasoning important to Kohlberg's sixth stage of moral development – still plays a part in the new synthesis theory of moral development, but it is of less significance than in Kohlberg's theory. Instead of learning to decipher between right and wrong through verbal reasoning, Haidt argues we learn through our emotional reactions to stimuli in childhood (Haidt 2007, p.998). In these emotional reactions, we learn to distinguish between good and bad. The process of moral reasoning occurs at a later stage, and usually concludes only to confirm the initial hypothesis based upon the emotional reaction. This process, in which emotional reactions take precedence over moral reasoning, continues throughout adult life, and is also noted in subjects whose brain injuries led them to perform unusual or illogical actions (Haidt 2007, pp.999-1000). Rather than express confusion at their own actions, these subjects would invent a narrative to justify the action. A moral justification for actions is always of importance to individuals, but humans are capable of finding moral reasons to justify almost any kind of behaviour. These moral reasons do not provide the basic motivation for our actions; instead, they follow behind our actions.

It is in the realm of affect where Haidt locates the roots of our actions and beliefs, though the roots go deeper than the emotional reaction itself. There are evolutionary reasons why people are predisposed to certain emotional reactions. A sense of morality contingent on emotional reactions allows for the formation of moral communities (Haidt 2007, p.1000). The moral community has a shared set of values that it expects its members to adhere to, along with a system of rewarding cooperation and punishing violators. As the moral community comes to realise that certain types of behaviour are beneficial to the group, the genes of its members will be modified to encourage these behaviours. Thus, people will be genetically predisposed to the values of the moral community to which they belong, with their emotional reactions to stimuli reflecting this evolutionary process. These genetic predispositions affect all aspects of our individual and social lives. As Haidt argues, "it is clear that we are prepared, neurologically, psychologically, and culturally, to link our consciousness, our emotions, and our motor movements with those of other people" (Haidt 2007, p.1001).

Christopher L. Suhler and Patricia Churchland have criticised Haidt’s approach to the study of moral development (Suhler & Churchland 2011). Innateness, they argue, is highly controversial within philosophy. Making too strong an argument about the innateness of a characteristic will lead it to being unapplicable in many situations. Too weak an argument leads to claims which are so broad and widely applicable that their explanatory power is negated. While Haidt is aware of the former problem, Suhler and Churchland argue he overlooks the latter (2011, p.2105). For Suhler and Churchland, Haidt's claims about the innateness of certain moral characteristics amount to nothing more than saying we are "prepared" to act in certain ways. In the same sense, we are prepared to ride bicycles, as many humans are well-equipped for the task from an early age. The task itself still has to be learned, however; we are not born knowing how to ride bicycles. Furthermore, Suhler and Churchland argue that Haidt's theory is not supported by the neurobiological data (Suhler & Churchland 2011, pp.2109-2111). Too little is known about the workings of the brain to be able to state that certain traits are innate. At this stage, they argue, it is overly ambitious to draw connections between neural activities and behaviour, and argue that these relationships are innate within the human condition.

Despite this critique, there are similarities between Haidt's and Churchland's approaches to moral development. Churchland also looks to neurobiological and evolutionary data to explain moral behaviour (Churchland 2011). Churchland links our biological nature with evolutionary survival; we have developed traits that better ensure our survival in the world. Our sense of morality is dependent on these biological traits (Churchland 2011, p.11). Primarily, these traits concern cooperation and care. Those who cooperate with others increase their chances of survival through receiving benefits from others in return (Churchland 2011, pp.63-94). The hormones oxytocin and arginine vasopressin are particularly important in encouraging us to care for each other. Churchland posits that high levels of these hormones lead individuals to form long-term attachments to others (Churchland 2011, pp.48-53), and to care for offspring. People with lower levels of these hormones struggle with these behaviours. From the possession of these traits, we are then prepared for social life (Churchland 2011, pp.118-162). We learn primarily to care for those close to us, and then to care for others beyond our closest social group. At this stage of Churchland's argument, there is a clear similarity with Haidt's approach. Churchland also argues for the importance of emotion and sociability in the development of our sense of morality. Rather than thinking of morality through our capacity for reason, learning to obey rules as deontologists posit, or thinking about the outcomes of actions as consequentialists argue, we develop our sense of morality through our intuitions (Churchland 2011, pp.163-190).

These theories do not provide an exhaustive account of the study of moral development in psychology. Other theorists such as B.F. Skinner and Freud (Freud 1962; Skinner 1971; Haidt 2008, p.67) have developed psychological theories of moral development. As Haidt notes, however, Freud’s theories are rarely discussed in contemporary study of moral development. Despite their epistemological and methodological differences, a common theme throughout theories of moral development is the importance of environment and sociability on the one hand, and emotion on the other. In Kohlberg, this is particularly evident at the conventional level of development, in which the individual’s primary concern is respect of the rules and values of the group to which they belong. For Haidt, it is the group that moulds the individual’s sense of morality, with evolutionary and biological factors feeding into this process. For Churchland, too, it is through the learning of social skills that we develop a sense of morality. The environment is the only concern for Skinner, who thought that human behaviour was reactive, and that it was the environment that dictated our thoughts and actions (Skinner 1971). Both Haidt and Churchland argue that the process of evolution leads to the development of emotions that determine our sense of morality, in conjunction with how our social group encourages their development. Emotion is of less importance for Kohlberg. However, empirical studies revealed that people do not all reach the final stage of moral development, in its Kantian formulation. Rather than through processes of moral reasoning leading to either deontological or consequentialist conclusions, it is the group to which we belong and our emotions that, in the main, determine our sense of morality.

In sum, Kohlberg's theory of moral development and Haidt's new synthesis provide the two main strands of knowledge in psychology on the origins of moral judgements. They do not stand in direct opposition to one another. Though there are differences in their approaches, there are also similarities. Both recognise that groups – whether families, communities, or nations – are of importance, and that the environment in which a person is raised shapes their moral development. However, whereas Kohlberg thought that the process of moral development was universal, with stage six being the result of any person's complete moral development, Haidt understands morality as being divisive. According to the new synthesis, different moral communities have different motivations guiding their development. A community at high risk of war would develop different moral values from a community that was peaceful and stable. The way in which a person's moral development occurs will be dependent on the community to which she belongs, and the values that this community holds.

3.6 Kantian Morality and Moral Communities

As political liberalism requires some kind of moral education in order for the individual to develop into a fully autonomous citizen, I assess the compatibility of a politically liberal moral education with theories of moral development. Following Dwight Boyd, I argue that there is congruence between the work of Kohlberg and Rawls. However, political liberalism is then subject to the same concerns revealed by empirical tests as Kohlberg's theory. The majority of people do not develop beyond the conventional level of moral development, meaning individuals respect the moral values of the group, but only appreciate these values insofar as they are of importance to the group. The person who acts as an aspiring moral philosopher, understanding the importance of moral values in the abstract, is a rarity. Haidt's findings pose further problems. Our moral development, according to this theory, is dependent on the moral community to which we belong; the moral community begins to shape our sense of morality even before birth. For the stability of political liberalism, these moral communities would need to be highly regulated to ensure a commitment was made to justice as fairness. The result of this would be to construct a system of moral education that prevented citizens from being psychologically motivated to endorse unreasonable doctrines.

Boyd has drawn connections between the work of Rawls and Kohlberg. As aforementioned, Snarey thought that Kohlberg's postconventional level of moral development reflected the philosophical ideals of Kant and Rawls. Kohlberg himself thought that stage six was comparable to the Kantian categorical imperative (Kohlberg and Hersh 1977, p.55). A person living in accord with such a sense of morality understands and acts from abstract moral principles, the basis of which would be universally applicable in any circumstance. The Kantian categorical imperative – on which a person acts as if the act was to become a universal moral law – is such an abstract moral principle. It is not a concrete command, as are the Ten Commandments. Kohlberg and Rawls both belong to the social contract tradition, and their philosophical ideals are in harmony, according to Boyd (2015, p.34). They both begin from the same premise; though in social relations there can be disagreement, there is a *right* that can act as an adjudicator (Boyd 2015, pp.37-38). Boyd argues that Kohlberg and Rawls both come to Kantian conclusions regarding the equal worth of all (Boyd 2015, pp.44-43). This is echoed by Rawls in the importance he ascribes to respect, and by Kohlberg in his understanding of sympathy, through which individuals come to respect others. According to both, individuals are capable of following rules, and they are self-determining agents (Boyd 2015, pp.48-52).

Rawls acknowledged that the political constructivism of justice as fairness had a Kantian basis (Rawls 1980, p.515). Both Rawls and Kant view the liberal values of equality and freedom as founded on respect, and its correlate, dignity,[3] according to Martha Nussbaum (2011a, p.2). Because of their inalienable dignity, humans must be respected as ends, never means, as embodied in Kant's categorical imperative. As Martha Nussbaum argues, this is why respect is so crucial for Rawls (Nussbaum 2011a, p.2). Through respect for others and our own self-respect, we recognise the necessity of political and social arrangements that promote freedom and equality. Furthermore, our freedom is restricted by the consideration of respect. In Kant, this is achieved through the recognition that our aims in life must be compatible with the autonomy of others (Anderson and Honneth 2005, p.127; Kant 1797). In Rawls, the same aim is achieved through reasonableness (Rawls 1980, p.530), which ensures our rationality, or self-interest, is restrained by ensuring its compatibility with the rationality of others. Full autonomy also recognises the constraints of reasonableness. A course of moral development leading to Rawlsian full autonomy would, then, come to largely Kantian conclusions.

Kohlberg thought that the stages of moral development were universal. If this were true, the moral development of any person, in any culture or society, would accord with the moral ideals of Kant and Rawls. Empirical evidence, however, revealed that Kohlberg's theory lacked universality. Most people's moral development ends at the conventional level; only a select few develop a sense of morality at the post-conventional level, where they view abstract moral principles as being important in their own right. As the majority of people only develop to the conventional level, the group to which they belong will be the primary arbiter of their sense of morality. People would not view justice as fairness as an important ideal in the abstract, but only insofar as obedience to the ideal was important for the group's stability. Values such as duty would take priority; upholding the social order would be of the utmost importance for the majority. For the citizens of political liberalism to become autonomous in the way Rawls intended it would be necessary for them to reach the post-conventional level. Once attained, they would be able to appreciate universal moral values for their own sake, endorsing principles through their capacity for

[3] Nussbaum notes that Rawls does not write about dignity, but she argues that the concept of dignity embodies the perspective on human nature he aims to defend.

moral reasoning. Through this capacity, they would be able to recognise the importance of justice as fairness for its own sake, and through reflection on differing conceptions of the good, decide on the development of their second moral power. Without the ability to do this, a person's sense of morality could not transcend the values of the moral community.

According to the new synthesis, however, the stages of moral development do not exist. Instead, we are primed for making moral decisions from birth, but rather than developing through a set of universal stages, development is hinged on the needs of our social group, or our moral community. This theory poses further challenges for political liberalism. If the theory were true, the moral communities in which we live would have a great deal of influence over our senses of morality. The way in which the two moral powers developed would be, to a large extent, dictated by our circumstances and surroundings. In order to ensure the first moral power developed to recognise justice as fairness, the moral communities would have to be strictly regulated. This would bring the state into tension with groups such as religious communities and families who home school. Moreover, as aspects of morality are genetic according to this theory, unreasonable values could enter society through birth. Public education would have to ensure that these unreasonable values were not developed. If education were left under the auspices of the private sphere, values at odds with justice as fairness would have the opportunity to develop, and the strength of their support could outweigh that of justice as fairness. The truth of the new synthesis would necessitate a robust public education system. Regulation of the private sphere to discourage the development of not only unreasonable beliefs, but also unreasonable genes, would be necessary for the stability of a well-ordered society.

A state that recognised that its stability and continued existence was dependent on ensuring the elimination of certain genes would appear to be highly authoritarian and illiberal. But if we assume the truth of the new synthesis, this would be the reality of all social groups: the family, public institutions, the nation, and the state. All of these groups prime their members toward certain dispositions, and in so doing, over time, ensure the development of some genes and the eradication of others. The state and well-ordered society of political liberalism would be no different. Through ensuring the realisation of psychological dispositions toward fairness and reasonableness, the state of political liberalism would be affecting our genetic heritage. If this process were implemented across the whole of society, in both the public and private sphere, the individual would become genetically primed to accept the values of political liberalism. This would undermine our motivation to endorse other doctrines alongside justice as fairness, as we would be emotionally predisposed to accept the principles of justice. In developing full autonomy, rational autonomy is undermined.

In addition to the problem of emotional predisposition – which is both necessary for an individual to become fully autonomous, and problematic for the diversity of beliefs – is the question of the good. For Rawls, the development of our second moral power is a matter of choice. We are to choose how to develop our conception of the good. If we are all to become fully autonomous citizens, however, developing to the final stage of moral development, a great demand is placed on our sense of morality. There are perhaps two ways the development of autonomy can occur in accord with theories of moral development. Either a society can be committed to ensuring citizens reach the final stage of moral development, where they can act from abstract moral principles, or ensure citizens are emotionally predisposed to accept justice as fairness. Each approach would, I argue, reduce the diversity of beliefs in a society. I now assess each in turn.

On the Kohlbergian scale of moral development, reaching the final stage of moral development, an individual learns to act from abstract moral principles. Empirically, only a small amount of people reach this stage of moral development; essentially those who study moral philosophy. If we are to attempt to enable everyone to reach this stage, then the whole of a society must be committed to the importance of moral philosophy. Citizens would need to view this as part of their conception of the good. That is, people would need to see the life of a moral philosopher as an ideal way of life. The outcome of this is necessarily perfectionist. Without this motivation, however, the existence of the fully autonomous citizen is something that can only be hoped for.

The alternative is to follow an approach which prioritises emotion over reason. Rather than developing to the post-conventional level of Kohlberg's theory of moral development, and understanding the importance of abstract moral values, most people remain at the conventional level of moral development. They take their moral ideals from the group to which they belong. However, Kohlberg underestimated the strength of the group's influence. Individuals do not endorse their group's ideals through their capacity for moral reasoning. Instead, as Haidt argues, they do so because they are emotionally predisposed to these ideals. The citizens of the politically liberal state would then be predisposed to the ideal of justice as fairness, at the expense of competing ideals. A system of moral education that inculcated a strong psychological motivation to endorse justice as fairness would preclude an individual from endorsing other doctrines. There is an implicit problem of comprehensiveness in any system of moral education, and the influence that the group has over the individual's sense of morality. For a citizen to become fully autonomous, however, some form of moral education would be necessary. Given the absurdly ambitious aim of developing a society inhabited solely by moral philosophers who act from abstract principles, it seems that ensuring citizens are emotionally predisposed to accept justice as fairness is the more appropriate way to ensure the development of full autonomy.

3.7 Conclusion

Political liberalism's approach to human nature is restricted by its need to avoid metaphysical commitments. Therefore, Rawls devised the political conception of the person, a concept that enables us to imagine certain basic features of human nature without reference to a comprehensive philosophical or religious perspective on the matter. Central to this conception are the two moral powers: the ability to form a sense of justice and a conception of the good. Through this conception, we come to think of ourselves as free: free to form a conception of the good, pursuing a rational plan of life in light of this conception; free to make claims on institutions; free to take responsibility for ends arising from our conception of the good. This line of thought undergirds Rawls' thought on autonomy. Through rational autonomy, an individual is able to determine her own conception of the good. In addition, the rationally autonomous citizen is able to decide on fair terms of cooperation. This leads to the acceptance of the principles of justice, in which fair terms of cooperation are embedded. Citizens are fully autonomous when they come to act *from* the principles of justice. They recognise that the principles of justice promote their self-interest as the principles allow an individual to advance her conception of the good. Because of this, individuals are motivated to endorse justice as fairness.

In order to realise the values of political liberalism – fairness, reasonableness, and a well-ordered society based on terms of fair cooperation – a person's moral powers have to develop in accordance with the principles of justice. This raises the problem of how this is to occur without the state imposing a perfectionist doctrine on future citizens through moral education. Rawls recognises that the political conception of justice must act as an educator. A state must impose certain values on to a society in order for it to function, excluding values that are antithetical. However, in the crucial stages of a person's moral development during childhood, imposing the values of political liberalism would have a strong psychological effect. The results of this psychological effect could potentially preclude a person from endorsing other religious or philosophical ideas. Though justice as fairness is devised so as to be compatible with other doctrines, if a person's moral development had been exclusively informed by liberal ideals, she may not be motivated to commit to other modes of thought.

Nevertheless, this may be a necessary course of action if we are to allow for the development of full autonomy. The stability of the society Rawls imagines is dependent on citizens who act as to uphold the principles of justice. Without citizens motivated to act in this way, political liberalism faces an existential threat. Rawls recognises that the psychological effects a society has on its citizens will reduce the diversity of beliefs within that society. He sees this as regrettable but necessary (Rawls 2005, p.200). In conflicts arising between the senses of rational and full autonomy, then, preference must be given to full autonomy.

The purpose of this chapter has been to assess how Rawls understands human nature and autonomy in political liberalism, in order that this understanding may be assessed against theories of free will in Chapter 5. In arguing that we must be held responsible for our conceptions of the good as rationally autonomous citizens, Rawls is tacitly committing to the truth of free will. If we are to be thought of as morally responsible, we must be held capable of freely choosing our own courses through life. On Rawls' understanding of rational autonomy, this is assumed to be the case.

Chapter 4 – Metaphysics and Science

4.1 Introduction

Following the explanation of the basic argument of *Political Liberalism*, and the role that Rawls' accounts of human nature and autonomy play within this argument, I now turn to the matter of metaphysics and science. Rawls thought that, as modern societies are already divided by many philosophical and religious doctrines, the justificatory reasons for constitutional essentials should not appeal to the truth of any one such doctrine, as this would be to risk deepening these divides. Instead, the realm of politics – processes of political deliberation, the functions of public institutions, constitutional matters – should be detached from wider philosophical questions, depending only on public reasons that all reasonable citizens can accept. Thus, metaphysics – which are, for Rawls, claims about universal truths – should be avoided throughout the course of political deliberation. In our justifications for the use of state power, we should not rely on the truth of one particular metaphysical doctrine.

In this chapter, I argue that Rawls was mistaken. This is partly due to the premises from which Rawls himself began. Rawls thought that there was a sense of self-respect within the human condition tied to the two moral powers. I argue that this in itself is a conception of the self containing metaphysical implications, as such a conception precludes other metaphysical conceptions of the self. Hence, Rawls does not remain entirely metaphysically neutral. Rawls' own account of human nature exists in tension with other doctrines within science, religion, and philosophy. Another problem with Rawls' attempt to restrict metaphysics from politics is that he argued that public reason should be restricted by the limits of well-established and non-controversial science. The relationship between science and metaphysics is close and complex. Through the methodological tools Rawls offers us, we cannot parse where the division between the two domains lies. Thus, I argue that if we are to consider the implications of science within public reason, we must also consider the metaphysical aspects of the science. We have reason, then, to consider neuroscience and its metaphysical implications within politics.

With the theoretical problems explained in Chapters 1 and 2, in this chapter, I establish the methodology that is used across subsequent chapters. Rawls' method of reflective equilibrium is used as a way of assessing theoretical claims against empirical data. I explain reflective equilibrium – including Rawls' original formulation and how it has been developed by other political theorists – before discussing its role in this thesis. This sets out the position I take in regard to the relationship between science and political theory.

The relationship between science and politics as Rawls understands it is assessed in 4.2. While Rawls allows for the inclusion of science within political deliberation, conditions limit what aspects of science we can consider. I argue against these conditions, arguing instead that, providing the implications of science do not challenge the basis of moral equality, we are justified in considering them within the political domain. In this section, I also explain reflective equilibrium, and establish the position I take with regard to how science and empirical data should be assessed against political principles. I then turn to the relationship between metaphysics and politics in 4.3. In arguing that we have two moral powers and a sense of self-respect, Rawls, I argue, is committed to a conception of human nature that does not cohere with many other theories within religion and philosophy. Following Jean Hampton and Joseph Raz who argue for the consideration of metaphysics, I argue that, as political liberalism is not entirely void of metaphysical considerations, we have reason for considering metaphysics within politics. Instead of drawing assumptions regarding human nature, we should ground how we think about human nature on an appreciation of the empirics. Thus, in 4.4, I argue that we should consider the metaphysical implications of science. It is on these grounds that I argue we should assess theories of human nature against the data in neuroscience.

4.2 The Place of Science within Political Liberalism

Public reason, Rawls argues, can be supported by well-known and non-controversial scientific knowledge. To lend support to the political principles we propose, we can appeal to such scientific knowledge. In this section, first, I offer an explanation of reflective equilibrium. I argue that if scientific theories are considered within reflective equilibrium, the implications of scientific knowledge must take precedence over political principles. While we might revise the political principles chosen in light of scientific knowledge, it seems unlikely we would wish to revise our understanding of science due to conflict with the political principles we find most convincing. Second, I argue that while this gives us reason to consider science as possessing authority within political deliberation, this authority must be placed within certain confines. The moral equality of all citizens, as understood from the perspective of the original position, should limit what sorts of claims can be made in light of scientific knowledge. Claims must not challenge moral equality between citizens. Third, I question the conditions Rawls placed on the consideration of scientific knowledge. Rawls argued that if science was to be considered within public reason, it must be well-established and not controversial. I present reasons for rejecting these conditions.

A note here about Rawls' method of reflective equilibrium is apposite. Reflective equilibrium is a method by which our considered judgements and moral principles are assessed against one another (Rawls 1971, pp.47-50). It is through a process of reflective equilibrium that Rawls arrives at the two principles of justice. *Judgements*, here, are our thoughts on issues at any level of generality, that we have arrived at in a clear frame of mind – we are not in a state of terror or inebriated, for instance – whereas *principles* are rules that guide deliberation in an enquiry (Rawls 1971, pp.46-50; Knight 2017, p.46). Where there is discord between a judgement and a principle, we revise either in an attempt to reach a state of equilibrium between judgements and principles. As Rawls notes (Rawls 1971, pp.49-50), a comprehensive survey of every principle relating to the subject under study, and all of the considered judgements that might arise while considering that principle, is impossible, but that if we consider the dominant principles from within moral philosophy that are familiar to us, this will move us closer to the ideal state of reflective equilibrium.

Though Rawls originally devised the concept of reflective equilibrium, it was further developed by Norman Daniels (1979). Daniels argued that reflective equilibrium should not only include principles and judgements, but also what he calls "background theories". While the main task of reflective equilibrium remains one of finding consistency between judgements and principles, background theories may alter the judgements we make. Background theories could be, for instance, conceptions of human nature or particular scientific theories that force us to further revise our considered judgements. To provide an example, if we found that the judgements we held in response to the principles guiding deliberation contradicted basic aspects of evolutionary theory, we would wish to revise our judgements rather than attempt to explain why evolutionary theory is incorrect. This may not always be the case; where a scientific theory is particularly controversial, we may choose to attribute greater weight to our moral and political judgements. For example, if empirical data appeared to show that there were biological differences between groups of people, we would still reject the judgements of those who argue against moral equality. As

Carl Knight notes (2017), it can be difficult to identify each component within reflective equilibrium – whether a particular aspect of an argument is a principle, a judgement, or a theory – but that this should not necessarily cause us a problem. The method of reflective equilibrium is a way for us to organise our thinking. For our purposes here, then, we are largely dealing with background theories within political liberalism. Conceptions of autonomy and the implications of neuroscience would be considered background theories within reflective equilibrium, which may force us to revise our judgements.

Considering science leads to further problems within reflective equilibrium. Knight states that it would be unreasonable to revise our understanding of science because it does not accord with our preferred political judgements (Knight 2017). We would not alter our understanding of the laws of gravity if it was somehow found that these laws contradicted certain aspects of our preferred political doctrine. Instead, we would revise the judgements we made in light of this doctrine, where these judgements sat in discord with such well-established science. In this sense, science possesses a higher level of authority over deliberation than do our moral convictions.

However, this is not always the case. Science, or at least the implications we draw from science, should not always possess a higher degree of authority than our moral convictions. For instance, a scientific theory on the psychological differences between men and women should not lead us to accept that there is not moral equality between the sexes. It is possible that scientific theories could be used to support fascistic ideas, for instance; ideas which would be rejected by parties in the original position. It should be this condition – the protection of moral equality between people – that determines whether we should consider the implications of science as possessing authority within political deliberation. This is not to say that would necessarily reject the truth of empirical data or a scientific theory on these grounds, but that this condition determines the appropriate implications we can draw from the science. Moreover, it seems unlikely empirical data or scientific theories would ever directly conflict with the concept of moral equality. While the implications of such data or theories may be used to support inegalitarian claims, we can reject these claims without making claims regarding the accuracy of the data or correctness of the theory. In sum, we should consider science as possessing a degree of authority within political deliberations, providing scientific knowledge is not used in an attempt to undermine moral equality, as such arguments would be rejected within the original position.

There is also a question of where it is appropriate to consider science. While Rawls argues that science can be considered within public reason, I argue it should be considered throughout the scheme of political liberalism. Congruence should be sought between the conditions of the original position and public reason. Suppose the implications of scientific knowledge were considered through public reason, but not in the original position. Public reason may reveal flaws in the conclusions reached in the original position. Thus, rather than risk the consequences of divergence between the original position and public reason, we should consider the implications of science across both. Scientific knowledge can, then, be used to support public reasons for the adoption of political principles (Rawls 2005, p.67). However, Rawls stated that this knowledge should be "well established and not controversial". Rawls only mentions these conditions briefly in *Political Liberalism* (see p.67 and p.224). There is little explanation offered as to what these conditions mean, or examples used to explain what aspects of science meet these conditions. However, we could imagine that the phenomena of gravity, evolutionary theory, atomic theory, or cell theory might fall into the category of well-established and non-controversial science. These theories have been accepted within the scientific community for generations, and most people should know something about the very basics of each theory through the course of their education, meaning they are well-known and established. There is also little disagreement over the general tenets of each theory – though there might be disagreement over the correct formulation, such as interpreting gravity through the theory of relativity or quantum theory – making such theories non-controversial, at least to scientists. Perhaps the theory of relativity or quantum mechanics may also qualify, though there is a deeper level of disagreement over the correct formulation of either theory, meaning appealing to specific aspects of one of these theories may prove controversial. If we are to appeal to such science, then, we should appeal only to the most general tenets of either.

There is also the question of who it is that finds the science controversial, as William Galston (1995, p.520) and Karin Jønch-Clausen and Klemens Kappel have noted (2016, p.7). Some religious and spiritual groups may question the validity of any of these theories, due to the science clashing with their theological or metaphysical perspective. Rawls is not clear on whose concerns about controversiality are to be respected. As Jønch-Clausen and Kappel argue (2016, pp.7-8), it is likely that Rawls was referring to the concerns of the scientific community when he spoke of controversiality – the science supporting public reason should not be highly controversial within this community – but this leaves us with the problem of what to do if the public disagree with the scientific community.

Presumably, though again Rawls does not explain his reasons for these conditions, making the acceptance of science within public reason conditional is to avoid a situation in which political principles are premised on highly contentious scientific knowledge. In the event of this knowledge being rejected, we would be forced to also reject the political principles. However, it could be the case that, whether or not political principles are knowingly premised on scientific theories, political principles come into conflict with science due to the progression of scientific knowledge revealing flaws in the assumptions undergirding these political principles. It would be irrelevant, then, whether or not political principles are premised on scientific theories. If we are to respect the validity of scientific knowledge, then we must acknowledge that political principles may come into conflict with such knowledge, and in the event of it doing so, we should revise the political principles. For instance, consider what Bertrand Russell states in regard to the state of scientific knowledge following advances in the seventeenth century:

> The first thing to note is the removal of all traces of animism from the laws of physics. The Greeks, though they did not say so explicitly, evidently considered the power of movement a sign of life . . . The soul of an animal, in Aristotle, has various functions, and one of them it to move the animal's body. The sun and planets, in Greek thinking, are apt to be gods, or at least regulated and moved by gods (Russell 1946, pp.493-494).

Following the discovery of the first law of motion, it was revealed that the external causes of motion were material. Suppose that, upon an animistic understanding of the laws of physics, we thought that there must be gods responsible for the laws of motion, and therefore considered it natural to codify respect for these gods within political principles. This codification may not be the result of considered judgements made knowingly in light of science, but the result of what is considered common sense. After Newton's discovery of the laws of motion, what was previously thought to be common sense could no longer be considered as such. That is, any argument for the necessary existence of gods could not rest on common sense appreciation of the laws of nature. It would no longer be a matter of common sense to codify respect for the gods within political principles once Newton had revealed the flaws in such assumptions. Hence, arguments could be made for altering these political principles. In such a case, it does not matter whether the political principle was knowingly premised on science. Its instability is revealed anyway through the progression of scientific knowledge.

Whatever the state of knowledge a political principle is devised in, however, there is the risk that untested assumptions are carried over into the principle. Understanding of what is common sense may alter as scientific knowledge progresses, revealing flaws in the assumptions. As what might be considered common sense alters, the sorts of political principles that appear sensible to adopt will need to adjust. Hence, whether or not political principles are knowingly premised on scientific knowledge, principles may conflict with such knowledge. Rather than seek to avoid the implications of controversial scientific knowledge, we should instead consider such implications through public reason. By bringing science and its implications into consideration, through a process of reflective equilibrium, the chances of error should be reduced.

Furthermore, the extent to which aspects of science are well-established and non-controversial is itself a political question. Michel Foucault claimed there was a relationship between truth and power in which the powerful can determine what is true (Foucault 1972). Whether or not we agree with Foucault, it can be admitted that what kind of scientific knowledge comes to be considered well-established and non-controversial depends on this relationship, at least to some extent. Through the basic structure of society, certain strands of scientific knowledge can be promoted while others are placed to one side. Governments can fund certain areas of scientific research, education systems can provide students with an understanding of some aspects of science, and the media can choose which areas of development in scientific knowledge to report, while other areas of science are ignored or suppressed. Those strands of scientific knowledge that are not promoted will neither become well-established nor non-controversial, due to the making of political decisions. To state that only those strands of scientific knowledge that are well-established and non-controversial can be considered within political deliberation would be to uphold the power the state has over the truth. Where science can reveal inconvenient truths, the state, appealing to Rawls' conditions, could justify the suppression of scientific knowledge from political consideration. This gives us reason to consider science within democratic processes, rather than allow the state to determine what sorts of scientific knowledge become well-established and non-controversial, and, in turn, which aspects of science can be considered within political deliberation.

Thus, I argue that we should reject the two conditions Rawls placed on the inclusion of science within political liberalism. Instead, the only condition for the inclusion of science should be that science is not used to defend the views of those who reject moral equality. There are two possible objections here, the second following from the first. First, the inclusion of all aspects of science would lead to instability. The political principles we endorsed would need to alter continually as the state of knowledge within science develops, leading to constitutional instability. Second, within this instability, there is no way of determining where the truth lies. There is no widely accepted position in the philosophy of science that can be appealed to in order to determine which aspects of science are appropriate to support political principles. Furthermore, the majority of citizens, who we can assume are not experts in science, have little ability to decide which is correct between two contrasting perspectives within the scientific community. Thus, the support which science was intended to lend political principles would be lost amidst confusion surrounding how to interpret data and theories within science. As Rawls states in his explanation of the burdens of judgement, often scientific and empirical data is too "conflicting and complex" for there to be common agreement on its interpretation (Rawls 2005, p.56).

These concerns should not trouble us, however. In response to the first objection, as stated earlier, whether or not political principles are premised on a knowing consideration of science, we may have to alter political principles as scientific knowledge progresses. Thus, knowingly committing to a position within science when devising political principles does not lead to constitutional essentials that are any less stable. In response to the second objection, I follow Hampton in arguing that disagreement is not problematic in itself; it does not matter, then, that most citizens are not experts in science and may disagree over how to interpret findings within science. As Hampton noted in her critique of Rawls' view on metaphysics, the main problem with disagreement is not one with metaphysics, but with "true believers" (Hampton 1989, p.812). The true believer is someone who is willing to do anything in order to advance the doctrine to which she is committed, will not listen to contrary opinions, and will shun those who object to her beliefs. Likewise, the problem here is not with science, but with those who would reject science without good reason due to their 'true belief'. It should be remembered that the limits of public reason apply primarily to constitutional essentials. Outside of these essentials – though Rawls argues public reason should still be respected (Rawls 2005, p.215) – citizens are free to appeal to any aspects of science that support their favoured policies. In these instances, citizens may feel marginalised if their perspective on science is disfavoured by the majority, but that does not mean that the views of these citizens are suppressed or not tolerated. Instead, these citizens are free to maintain a belief in the correctness of their interpretation of the science, arguing the case for this interpretation, and attempting to garner support. Returning to public reason and the attempt to establish constitutional essentials, the same rules should apply. There is no need to make the consideration of science conditional. Using findings within science to support a constitutional essential is not to be intolerant to those who reject these findings. Furthermore, to place conditions on the consideration of science is to risk rejecting matters of great urgency. For instance, a finding on climate change may be rejected from political deliberation due to its not being well-established. The consequences of ignoring this science could be a series of disasters making the state uninhabitable. From the perspective of the original position, then, parties would recognise the need to consider science within public reason due to its protecting of citizens' interests.

If we are to consider science within political liberalism, we should, then, consider it as possessing authority.[1] Though some may reject the empirical data, if we are to consider science as possessing authority within political deliberation, we may discount those who reject scientific knowledge as being unreasonable. Rejection of scientific knowledge is not reasonable disagreement; it reveals an unwillingness to recognise the burdens of judgement (Rawls 2005, pp.54-57). This claim may seem to be at odds with Rawls' original explication of the burdens of judgement. On

[1] Of course, this leaves us with the problem of who is to determine the quality of the scientific data if this data is to hold such authority. Elizabeth Anderson has offered a solution to this problem, devising a hierarchy of expertise between the layperson and the leading scientists in their field (Anderson 2011, pp.146-147). Anderson also proposes democratic reforms to allow for better dissemination of scientific knowledge, enabling the lay person to become informed. Whatever aspects of science are incorporated within the political domain, the question will remain of who is to determine the most appropriate interpretation of the science. Anderson's solution is a sensible one.

Rawls' view, as we are to respect the views of other reasonable people, not attempting to use the power of the state to suppress their views, we should accept that empirical and scientific evidence can be complex, and respect the judgements people make in light of this evidence. The power of the state should not be used to suppress the views of alternative interpretations of the evidence. However, if a person wishes to reject the scientific evidence because she finds it unappealing, despite making no attempt to offer an alternative explanation of the evidence, we should not consider this reasonable disagreement. Thus, a person who wishes to prevent the state from using force to mitigate the effects of climate change, though who does not offer a plausible alternative account of the data, should not be considered as reasonable. The problems of disagreement arise not from science and metaphysics, but, as Hampton argued, from the true believers who refuse to listen to reasonable arguments.

Removing the conditions stipulating which aspects of science can be considered means that we are free to appeal to all aspects of science in the use of public reason. As I argue in later chapters, considering the implications of neuroscience should lead us to revise how we think about autonomy. Nevertheless, the proponent of political liberalism may object to its inclusion on the grounds of its controversiality and its metaphysical implications. I have explained why we should not consider controversiality a problem, while I address the problem of metaphysics in 4.3 and 4.4.

To sum up, through a process of reflective equilibrium, we should consider science as holding a degree of authority within political liberalism. This authority should be tempered, however. Scientific knowledge should not be used to challenge the notion of moral equality between citizens. Rawls also thought that science should only be considered if it was knowledge that was well-established and not controversial. I have argued against adopting these conditions. There are two reasons for this. First, while science may not have been knowingly considered within the deliberative processes behind the devising of political principles, it may nevertheless have indirectly influenced these processes. What sorts of claims are considered common sense can be influenced by the current state of knowledge within science. Thus, the political principles we adopt may be influenced by scientific knowledge unknowingly. Rather than restrict the consideration of controversial scientific knowledge from political deliberation, we should allow it to inform deliberation in order to understand the sorts of assumptions that undergird political principles. Second, whether or not an aspect of science is well-established and not controversial is itself a political question. To reject scientific knowledge from political deliberation on these grounds is to increase the extent to which the state can control the development of knowledge and hide inconvenient truths. Thus, we should consider science within political liberalism; those who reject scientific evidence without an alternative explanation of the evidence can be considered unreasonable. However, this is not to say that those who reject science are not to be tolerated.

4.3 Political Liberalism and Metaphysics

Metaphysics, for Rawls, belongs to the domain of philosophical doctrines that should not be considered within the political realm. The sorts of truth claims that are the subject of metaphysics relate to the ways in which individual citizens may conceptualise the good life. Committing to a certain metaphysical doctrine within political deliberations over constitutional essentials would deepen divisions between citizens, as those who object to such a metaphysical claim would likely feel that the state is forcing them to obey a constitution premised on ideas they think untrue. In this section, I explain how Rawls attempted to remain neutral between metaphysical claims in politics, before examining how such arguments have been extended in the work of Rawlsian thinkers. I argue that Rawlsians are mistaken in thinking that they are entirely neutral in matters of metaphysics. In the conception of human nature they endorse, there are metaphysical implications preventing us from accepting the truth of other perspectives on human nature.

To begin, an explanation of metaphysics is necessary. The meaning of metaphysics is often unclear, and there is no universally accepted definition of the term (Van Inwagen 2018, p.1). For our purposes here, we will assess what Rawls understood metaphysics to be. As Jean Hampton notes (1989, pp.794-795), Rawls does not offer a precise definition of what he means by metaphysics. His loose definition of the term will inform our discussion. The closest Rawls comes to offering a precise definition of metaphysics is in a footnote in his reply to Jurgen Habermas. Rawls says here that metaphysics is a:

> general account of what there is, including fundamental, fully general statements – for example, the statements 'every event has a cause' and 'all

> events occur in space and time', or can be related thereto. So viewed, W.V. Quine also is a metaphysician. To deny certain metaphysical doctrines is to assert another such doctrine (Rawls 2005, p.379).

It is the final sentence here that is of especial importance. The problem with metaphysical claims, for Rawls, is their uniqueness; if a metaphysical truth is espoused, alternative truths are denied. In his essay 'Justice as Fairness: Political not Metaphysical', Rawls explains that his conception of justice is political rather than metaphysical because it provides a basis for political agreement, rather than a conception of justice that is considered to be true (Rawls 1985, pp.230-231). Expanding on this, Rawls claims the way in which he imagines citizens coming to this agreement as free and equal persons is not dependent on any metaphysical conception of the self. Though Rawls does not offer an explicit definition of metaphysics, his understanding of the term centres on universal truth claims. As Rawls puts it, "I should like to avoid, for example, claims to universal truth, or claims about the essential nature and identity of persons" (Rawls 1985, p.230). Thus, a conception of justice would be metaphysical if a person argued it was *the* true theory of justice. This true theory of justice would preclude all other theories of justice. Committing to it would prevent a person from endorsing an alternative conception of justice. Likewise, if someone were to argue that a theory of the self was *the* true understanding of the self, it would be a metaphysical perspective on the self, rather than a political conception of the self (a conception of the self that all people can agree is reasonable). These sorts of truth claims are metaphysical. While Rawls does not provide a substantive definition of the type of truth that undergirds metaphysics, the problem of a metaphysical truth is its uniqueness; that endorsing such a truth would negate alternative metaphysical perspectives.

Of course, this leaves certain problems unsolved. What types of truth claims are to be considered metaphysical, as opposed to scientific or empirical? Hampton argues that Rawls' thought on metaphysics is Hobbesian rather than positivistic (Hampton 1989, pp.794-795). That is, it is not merely non-empirical claims that we can dismiss without further consideration, but claims for which an "incontrovertible demonstration is not possible". On this definition, we can delineate between the scientific and the metaphysical. The metaphysical is that which cannot be demonstrated. However, this is not sufficient. As Raz claims (1990), if it can be shown that a theory can achieve its aims in practice, then we can consider it true. In this case, we have empirical proof of a theoretical claim, and can, then, consider it true. It would seem odd, however, to consider a theory of justice as a scientific rather than a metaphysical truth. That Rawls considers it *possible* for a theory of justice to be metaphysical, and claims made in regard to justice are demonstrable, suggests that Hampton's explanation of Rawls' thought on metaphysics fails.

To gain a clearer understanding of what metaphysics means for Rawls, it is better to consider the central problem of metaphysics for politics as Rawls sees it. The problem, as mentioned earlier, is that of exclusivity. Believing something to be universally true precludes any alternative perspective on that truth. A person who believed a theory of justice to be universally true would not countenance any other theory of justice. Though aspects of a theory of justice are demonstrable, its universal and exclusive truth cannot be demonstrated. Thus, Hampton's explanation succeeds if we think about the exclusivity of universal truths. This suggests, then, that the problem for Rawls is not one of metaphysics but of *belief*, as Hampton also argues (Hampton 1989, p.812). A scientist may believe her scientific theory to be universally true. Again, aspects of a scientific theory can be tested and proved false, but this may not challenge the belief. A person may continue to believe in the truth of a theory despite the availability of contrary evidence. While we may have a better understanding of what the problem of metaphysics is for Rawls, it is still not clear whether scientific truths are different from metaphysical truths if the truth claim is intended to be universal. More on this is said in 4.4.

For Rawls, metaphysics belong to what he calls the "background culture" of society (Rawls 2005, p.14). The background culture contains all of the ideas and beliefs people hold as they go about their everyday lives. While this culture is social – people form institutions on its basis, and interact with others with regard to the values they hold as a result of this culture – it is not political. We should not draw on aspects of the background culture in order to support our political principles. To do so would be to argue for comprehensive, rather than political, liberalism. Such a liberal doctrine contains views not only on what is just, but also on what is good and true, and guides all aspects of human behaviour on this basis. Political liberalism, on the other hand, confines itself only to what is just. The task of determining what is good and true is left for citizens to decide for themselves. From the constitutional essentials determined through political liberalism, we cannot necessarily determine what is good or true, and so – outside of the realm of what is just – human behaviour ought not to be guided by this doctrine. Providing that people act in accordance with what is just, how they act and think in relation to what is good and true is for them to decide.

Thus, citizens living in a politically liberal society should not expect their favoured metaphysical doctrine to be enforced by the state over other citizens. Given that Rawls thought that public reason should be restricted by certain aspects of scientific knowledge, however, it is not necessarily the case that political liberalism abstains entirely from making commitments to the truth. Quong's response to this problem is to determine what kinds of truths political liberalism wishes to avoid. Arguing here against Raz's claim that justice cannot be independent of the truth (Raz 1990), Quong distinguishes two types of truth: mundane and metaphysical (Quong 2011, pp.226-229). I focus here on Quong's interpretation of Raz in order to convey Quong's argument:

1. Rawls's theory aims at stability and social unity.
2. Aiming at stability and social unity explains the need for an overlapping consensus.
3. But we need to know why we should aim at stability and social unity.
4. The reason must be that these goals are what a true or valid theory of justice should do (according to Rawls).
5. But this means that Rawls's theory must accept certain claims as true, namely, that social unity and stability are of great value and should be the focus of a theory of justice.
6. Therefore Rawls's theory cannot successfully abstain from certain truth claims (Quong 2011, p.227).

Thus runs Raz's line of argument, according to Quong. If 6 is correct, and political liberalism cannot avoid truth claims, it is only the mundane type of truth claim, according to Quong (2011, p.228). That is, it is mundane rather than metaphysical. We may say that it is inherently true that social unity and stability are of great value to a society, but this sort of truth claim does not commit us to any one metaphysical doctrine, at the expense of opposing metaphysical positions. A Kantian, a Hindu, and a utilitarian could all agree that it is true that social unity and stability are important, despite their respective metaphysical perspectives being irreconcilable. Whether we say that justice as fairness is true, or only that it is reasonable, it remains the case that we can accept the precepts of justice as fairness without adjusting the metaphysical doctrines to which we are committed.

Though Quong, I argue, is partly correct here – justice as fairness in itself is compatible with various metaphysical perspectives – some of the ideas Rawls uses to support justice as fairness are not entirely void of metaphysical implications. This is particularly true of the way in which Rawls thinks about self-respect and the moral powers. Assuming that there is a sense of self-respect and two moral powers within the human condition – and that this provides the basis of our motivation in life, as it undergirds our moral psychology – precludes many other metaphysical perspectives on human nature. For instance, if we thought, as Locke did (1690), that the mind begins as a *tabula rasa* – that knowledge is dependent on experience, and the mind at birth, having no experience, can contain no knowledge – then we could not assume also that self-respect and the two moral powers are inherent within the individual. Such attributes of the human condition would be dependent on the knowledge produced by certain social circumstances. For instance, whether a sense of justice was maintained within a society, or whether it was considered appropriate to develop a sense of self-respect in response to personal success. On Lockean empiricism, we could assume nothing in regard to the human condition from the perspective of the original position. Or if we thought that humans were naturally egoistic, with no sense of morality other than that of self-preservation, as Hobbes perhaps did (1651), the notion of self-respect is entirely at odds with our account of human nature. The individual of Hobbes' state of nature could not be thought of as increasing her self-respect through realising her moral powers, nor would Hobbes imagine that this became the focus of human life under the sovereign. Freud thought that unconscious drives, primarily concerned with sex and death, were what motivated human behaviour (Freud 1920). The death drive in particular has a self-destructive aim. Any kind of moral power we might have would perform a secondary function to these primary drives. Freud would see that we are primarily focused on what derives from these unconscious drives; our success in realising our moral powers, or that of anyone else, is ultimately irrelevant in terms of our overriding motivations. This is a short summary of three ways of thinking about human nature that are inconsistent with Rawls' idea of self-respect and the moral powers as the basis of motivation. From the history of religion and philosophy, we could draw many others. Rawls' view of human nature precludes any other metaphysical perspective that does not find humans to be inherently capable of social cooperation due to their innate capacity for a sense of morality.

While justice as fairness may indeed be committed to nothing more than the mundane sort of truth, the same cannot be said of the way in which Rawls thinks about human nature. If we endorse Rawls' conception of human nature, we will be precluded from accepting many other philosophical conceptions of human nature. Thus, as I argue in 4.4, rather than base our understanding of human nature on such assumptions, we should instead appeal to empirical data. This allows for a more stable basis on which to form a conception of human nature. Furthermore, in thinking about human nature, remaining neutral in regard to metaphysics is not possible. Our understanding of human nature will come into conflict with other ways of thinking about human nature.

As Hampton argued (1989), considering metaphysics within the scheme of political liberalism lends the deliberations greater depth. Joseph Raz also argues for the need to consider metaphysics within political deliberation (Raz 1990). Raz argues that justice and truth cannot be independent of one another (Raz 1990, p.15). If a theory of justice attains the values which it was established to realise – in this case, fairness – then we can consider that theory to be a true theory of justice. The mere achievement of its aims means that justice cannot be independent of the truth. Jean Hampton also worries about the relationship between justice and truth. Disallowing metaphysical reasoning within political liberalism could lead to a situation in which political leaders and citizens use emotional rhetoric to support their principles, rather than attempting to establish the truth of their principles (Hampton 1989, p.807). Opening up political liberalism to deliberation on metaphysics allows for better grounds on which to debate certain issues. Controversial matters over which there is no consensus would sometimes be better settled on metaphysical grounds. To use Hampton's example (1989, p.810), a debate on the legality of pornography could not be determined through appealing to the principles of justice alone, or to questions of reasonableness. Those who argue pornography violates dignity and should be banned, and those who argue its being banned is a violation of freedom of speech are both being reasonable, in the Rawlsian sense. Our stance on such an issue is dependent on what we think is metaphysically and morally true; our understanding of human nature and dignity. Enabling the inclusion of metaphysics within public reason allows us to incorporate deeper reasoning to support our stance.

As noted in 4.2, Hampton stresses that this does not necessarily promote intolerance. While Rawls wished to keep metaphysics out of politics to avoid deep epistemic disagreements being transposed into the political realm, the presence of metaphysics is not an indication of intolerance in itself. Relying on metaphysical reasoning to support the principles we endorse is not to say that we think our opponents should be prevented from endorsing opposing metaphysical views.

As political liberalism fails to remain entirely neutral in metaphysical matters regarding the nature of the self, we should, I argue, resolve matters of metaphysics within political liberalism. Avoiding metaphysics within politics entirely is neither possible nor desirable. Where metaphysical issues arise, such as the nature of the self, rather than attempt to devise a conception of the self that is devoid of metaphysical commitments, we should instead attempt to brings judgements on human nature into accord with empirical data within science, seeking to attain reflective equilibrium.

Finally, there is the matter of what aspects of metaphysics are to be incorporated; whether any metaphysical perspective should be considered as grounds for a legitimate public reason, or only certain kinds of metaphysical claims. For instance, a person may wish to appeal to her faith in God, her understanding of the meaning of life, or her belief in the immortality of the soul in order to support her claims within public reason. Hampton's solution here is Socratic philosophising (1989, p.812-814). A search for truth through Socratic philosophising is quite different from attempting to impose a belief system on another. Through Socratic philosophising, we agree to respect the views of others in attempting to establish the truth. Our aim is not to simply impose what we regard as true on the rest of society. Whether we are attempting to reach political conclusions by Rawls' method of public reason or Hampton's proposed Socratic philosophising, the problem, as aforementioned, is the true believer. The true believer has no respect of alternative ideas to her own, as she wishes only to impose one belief system on the whole society. This is a problem regardless of the method of political deliberation. This will be further explored in Chapters 8 and 9. For our purposes here, Hampton's method of Socratic philosophising can be used to determine what sorts of metaphysical claims can be considered within politics. Where the implications of science – or other assessments of the truth relevant within political deliberation – lead us to considerations of metaphysics, the truth claims arising can be tested against opposing conceptions of the truth. Thus, we are not simply appealing to what we believe to be metaphysically true, but attempting to sharpen our understanding of the truth. Commitments to preconceived metaphysical beliefs are thus abandoned for the purposes of political deliberation.

To conclude, Rawls argues that as metaphysics are to be placed aside from the political domain, in political considerations, we should not draw on a metaphysical conception of the self. However, I have argued that Rawls fails to remain entirely metaphysically neutral in his account of the self, at least in regard to a Rawlsian understanding of metaphysics. Whether we say that the political conception of the person is a reasonable way of thinking about the self, or the true way of thinking about the self, should we accept this conception, we are precluded from accepting many other conceptions of the self drawn from the history of religion and philosophy. Thus, rather than base our understanding of the self on general assumptions made about human nature, we should instead look to the empirical evidence, and base any claims made about the self on an understanding of the empirics.

4.4 Why the Metaphysical Implications of Neuroscience Should be Considered

Due to the absence of metaphysics from political liberalism, free will is not a subject Rawls would have considered appropriate to deliberate in the original position. In this section, I argue that Rawls was mistaken. Metaphysics and science are entwined. As Martha Nussbaum notes, Rawls left it to future students of political theory to determine the boundary between the two (Nussbaum 2011a, pp.7-8). If science is to be considered, I argue the metaphysical assumptions underlying scientific knowledge need to be unpacked if we are to understand the implications of this knowledge. My aim is not to develop a theory of metaphysics, or to argue in favour of a particular approach to the philosophy of science, but to show that the methodological tools Rawls offers us do not allow us to determine where science ends and metaphysics begins. I argue that if, as Rawls claims, science should be considered within political liberalism, then there is a need to consider metaphysics. This discussion supports the four reasons that are then presented for considering neuroscience within politics. First, we can cannot distinguish metaphysics from science on Rawlsian grounds. Second, to consider the metaphysical implications of science is to gain a better understanding of what science tells us. Third, there are already metaphysical implications within political liberalism due to Rawls’ conception of the person. Rather than allow for the inclusion of this conception of human nature based only on assumptions made regarding human nature, we should test these assumptions against the data in neuroscience. Fourth, to consider these matters is not necessarily to be intolerant of alternative perspectives. Through the consideration of neuroscience, we gain a better understanding of human nature and its relationship to politics.

There is a blurred line of division between science and metaphysics. Craig Callender states that philosophers of science are troubled by what aspects of metaphysics to incorporate within their work as “metaphysics is deeply infused within and important to science” (Callender 2011, p.33). If public reason is to be supported by well-known scientific knowledge, the limits of this knowledge – what is strictly scientific as opposed to metaphysical – would also need to be understood, yet Rawls does not offer us the tools to determine this limit. While Karl Popper (1963) and Rudolf Carnap (1963) attempted to draw a sharp distinction between the domains of metaphysics and science, this distinction has become of less importance to some working in contemporary metaphysics and science (Callender 2011; Mumford & Tugby 2013). This is not to say that there are no longer dividing lines between the two domains, but that those such as Callender, along with Mumford and Tugby, recognise that each domain must inform the other. As Mumford and Tugby posit, scientific knowledge requires an ordered world. In a world that suffered from complete disorder, scientific knowledge could tell us nothing; there could be no way of applying this knowledge within this disorder. Within an ordered world, we can apply scientific knowledge, as its hypotheses are predictable and testable. The systems themselves that are constitutive of this order, however, are not knowable through scientific knowledge alone. To understand these systems, we need to look to metaphysics. Through combining scientific knowledge with metaphysical analysis, we can not only understand scientific data, but also how this data fits into the order of the world in which it exists, or at least a perspective on this order.

Determining what is scientific as opposed to metaphysical is, though, not a straightforward task. Perhaps a simple way of thinking about the distinction is to regard metaphysics as concerning abstract and speculative claims, and science as concerning what is empirical and testable. As Callender notes (2011, pp.10-11), however, this does not suffice. A scientific theory such as superstring theory contains claims that are speculative. In addition, the proponent of superstring theory posits that extended simples – simples being objects with no proper parts – are the building blocks of the universe, an idea some metaphysicians would argue is metaphysically impossible, as simples cannot be spatially extended. The relationship between the two domains is thus complex and far from clear. While recent work on the relationship acknowledges that metaphysics cannot simply be stripped away from science, there is no common agreement on how the one should relate to the other. Ross et al (2007) argue that contemporary metaphysics should consider science only as it is, discarding a priori intuitions and common-sense ideas uninformed by science.

Against this, Anderson and Becker Arenhardt (2016) argue that intuitions are important, and science presupposes the truth of certain metaphysical intuitions. Science and metaphysics are complexly related and there is no simple way of determining the boundaries between the two, or how they should inform one another.

If we are to consider science within public reason, but not metaphysics, we must know where the dividing lines between the two domains exists. Rawls does not offer us the ability to determine this through his methodological tools alone. The original position cannot solve this problem as deliberations are guided by the parties' interpretation of the citizens' interests, and we can assume that the interests of citizens are in no way dependent on the distinction between science and metaphysics. It would be of no benefit to citizens to define metaphysics in a particular way or science in another. Public reason also fails to offer a solution. We are to keep metaphysical considerations out of public reason; determining a distinction between what is metaphysical as opposed to scientific is itself a metaphysical question. Perhaps this is also why his definition of metaphysics is loose; to commit to a certain picture of metaphysics is, in itself, to commit to a metaphysical truth. A possible solution for Rawls exists in the demonstrability of a truth. If a truth can be demonstrated, then we can consider this truth within politics. Truths that are not demonstrable – a belief in God, the afterlife, or the meaning of life – are what should be avoided within politics, whether we consider these truths metaphysical or scientific. As Hampton noted, however, the problem here is one of belief, not truth. We have reason to exclude indemonstrable beliefs from politics – as Hampton argued, it is the true believer who is the problem for politics, rather than the truth – but we do not have reason to exclude truths. Using the implications of science and metaphysics to support our public reasons offers them a greater depth.

Though we can support public reasons with the conclusions of science, without metaphysical analysis of science, certain knowledge is of no value. This is not true of all scientific knowledge. For instance, we could understand that carbon emissions cause global warming through looking at the scientific evidence, with no need to unpack the metaphysical assumptions underlying the evidence. Science, here, has only practical implications; there is no need to question the metaphysical basis of the work. This is not the case, however, when scientific knowledge is considered in relation to other matters. Consider how scientific knowledge might be relevant to the legality of abortion at certain stages of pregnancy. We may wish to look to the scientific data in order to determine when certain physical processes begin – the heartbeat, the development of the central nervous system – in order to determine where life, consciousness, or the ability to feel pain begin. Scientific data relating to when such phenomena begin between conception and birth can tell us little without a metaphysical perspective on the meaning of these concepts. From the scientific data, we can learn about the beginnings of various physical processes, but not what these processes signal in relation to our understanding of life. Why the beginning of certain physical processes should be considered as the beginning of life, or personhood, is a metaphysical, rather than scientific, question. While scientists might make claims regarding what the data tells us, when they do so, there are certain metaphysical assumptions underlying their claims that are unacknowledged. As Peter Van Inwagen argues (2018, pp.7-8), when the scientist Carl Sagan asserts that the world and the physical universe are identical, he is making a metaphysical claim, not a scientific one.

Such issues raise metaphysical questions that need to be resolved metaphysically, not scientifically. Hume claimed that what ought to be cannot be inferred from what is (1740), yet there is a deeper problem here: the metaphysical *is* and the scientific *is* are not synonymous. The science alone can tell us about what there is scientifically, but not metaphysically. Empirical data without metaphysical analysis does not tell us how science should inform our perception of the truth. A scientific theory may tell us more about our perception of the truth, yet there will be certain metaphysical assumptions underlying this theory. Rather than attempt to include science within political liberalism while excluding metaphysics – a move which is impossible without defining each term, which Rawls does not allow us to do – we should include both. This allows us to consider the metaphysics that undergird scientific theories, and the metaphysical implications following from science and empirical data. From this consideration, we gain a better understanding of the science, as those such as Mumford and Tugby argue, while also allowing us to think about how the conclusions of science can inform our lives. Unpacking these metaphysical considerations allows us to draw conclusions on matters such as abortion.

If we allow for the consideration of the metaphysical implications of science, we can consider neuroscience. The political conception of the person can be assessed against the empirical evidence. From the metaphysical implications arising from this evidence, we can deepen our understanding of human nature, considering whether this conception of human nature coheres with our political principles. Due to the implications it may hold for how we think about autonomy, we should consider the empirical data relating to free will.

Because of the authority we grant science within public reason, where conceptions of autonomy do not cohere with the empirical data, we should revise how we think about autonomy. This is also true for other political concepts which can be related to human nature: moral and legal responsibility, political decision-making, or our sense of morality. Through assessing such concepts against the empirical data, we can better appreciate how political principles cohere with the implications arising from human nature. Though there are other aspects of human nature that are of political importance, it is the relationship between free will and autonomy that is considered here.

The subject of free will incorporates both metaphysical and empirical aspects, both of which, I argue, should be considered. Philosophical discussion of free will has tended to examine its metaphysical aspects, while neuroscience looks at free will from an empirical perspective. Empirical data necessarily leads us to metaphysical considerations, however. Without considering the metaphysics, empirical data alone tells us little about free will, as we are left without a metaphysical understanding of the meaning of free will.

Metaphysical explanations tell us how the empirical data challenges free will. Hence, whether we think there are implications from science for free will and moral responsibility depends on our metaphysical commitments, whether explicit or implicit. Libet's reading of the data is premised on the truth of mind-body dualism.[2] On this premise, we have free will due to our conscious will being independent of neural activity, the mind and body being two distinct entities. The data obtained from Libet's experiment challenges this conception of free will as it suggests unconscious neural activity has a large role to play in the process of human action (though this conception of free will is rescued by Libet's insistence on the indeterminism of the veto function). Without understanding this premise, we can learn only about a series of physical processes. We cannot draw conclusions regarding what these processes mean for whether or not we have free will. If we are to consider neuroscience, the metaphysical implications of neuroscience should also be considered.

[2] See Libet (1999) and Dennett (2003, pp.232-236). This is further explained across Chapters 5, 6, and 7. As Libet notes (2006), this is not necessarily to say that any part of the mind-body relationship needs to be immaterial, but that what determines the actions we choose to make needs to be independent of other causal factors.

Rawls would presumably respond here by saying that while we are justified in considering the data within Libet's experiments, providing it is well-established and non-controversial, we are not justified in considering its metaphysical implications. This response fails, I argue, due to four reasons. First, with the methodologies of political liberalism alone, we cannot know what is scientific as opposed to metaphysical. While Libet was a scientist, many of his claims were metaphysical. Without a clear distinction between the two, we cannot determine which claims belong to which domain. To commit to the truth of a distinction would be, however, to commit to a metaphysical claim. Second, from considering the metaphysics of science, we gain a deeper understanding of the data and theories within science. It would be fruitless to consider certain aspects of science without assessing the underlying metaphysics. Third, in his way of thinking about the self, Rawls commits to a picture of human nature that rules out other ways of thinking about the subject. Instead of committing to such a conception, we should test assumptions regarding human nature against the empirical data. To do this will involve assessing of metaphysics alongside the data. From this assessment, however, we gain an understanding of human nature that withstands scrutiny from an empirical perspective. Such a conception of human nature cannot be rejected without an explanation of why either the data is incorrect, or the analysis of the data fails. Fourth, it is not the truth of the conception of human nature that is problematic, as Hampton argued. It is indemonstrable beliefs to which people are committed, and the attitude the person holds towards this belief. None of this is necessarily true in the case of neuroscience or its metaphysical implications.

On these grounds, we can include the implications of neuroscience within the scheme of political liberalism. Theories of human nature can, then, be tested against the data within neuroscience, along with its philosophical implications, in order to devise a theory that coheres with the data. Such a theory will be cognisant of the problems for liberalism arising from human nature. Rather than assume a theory of human nature that sits well within liberal political theory, this theory is informed instead by the empirical data, assessing the relationship between human nature and liberalism. As I argue in Chapters 8 and 9, this will lead us to revise how we think about rational and full autonomy.

To conclude, if scientific knowledge is to be considered within public reason, metaphysics should also be included. The line between science and metaphysics is often blurred. Political liberalism does not offer us the tools we need to determine the boundaries of each domain. Furthermore, considering the metaphysical perspectives that lie behind scientific knowledge assists us in understanding the implications of this knowledge. Rather than consider science and metaphysics as two separate and independent domains, we should assess both together. This is reason, then, to consider the implications of neuroscience within political liberalism. This allows us to assess whether our way of thinking about human nature coheres with the empirical evidence.

4.5 Conclusion

I have attempted here to define the place of science and metaphysics within political liberalism. While Rawls claimed certain aspects of science should be considered through public reason, I have argued for a much wider consideration of science and its implications throughout political liberalism. While Rawls placed conditions on what aspects of science to include – those of its being well-established and non-controversial – I argue these conditions should be discarded. On these grounds, within certain limits, we can consider all aspects of science within political deliberation.

If we are to consider science, we should consider neuroscience. Rawls draws on several assumptions regarding human nature in his political conception of the person. Rather than deriving our understanding of human nature from assumptions, we should instead assess such assumptions against empirical evidence. There is, then, a need to consider psychology and neuroscience. This is relevant in considerations of moral development, how we come to hold moral responsibility, and how human behaviour relates to the structure of society or the legitimate use of force.

While it might be argued that these are matters of metaphysics rather than science, and as such have no place within political liberalism, I have argued that the distinction between what is metaphysical as opposed to scientific often is not clear. Rawlsian methodologies also do not offer us a way of making this distinction. Instead of attempting to draw this distinction, we should instead consider the metaphysical implications of science. Hampton's method of Socratic philosophising offers us a way of determining what sorts of truths are appropriate to consider. This is not a matter of appealing to beliefs and attempting to use the power of the state to suppress alternative beliefs. Instead, it is a method through which we can attempt to gain a better understanding of matters, while respecting opposing beliefs. Thus, through such a method, we can consider what science tells us about free will, bringing such considerations to bear within political theory.

Chapter 5 – Free Will and Political Liberalism

5.1 Introduction

Political liberalism requires a particular liberal subject. The main requirement of this subject – as identified in Chapter 3 – is that she is fully autonomous. That is, she has developed her second moral power as to recognise justice as fairness, acting in accord with the principles of justice. At the same time, we respect a person's right to choose; we respect her ability as a rationally autonomous citizen to determine her own good. In this chapter, I argue that Rawls' conception of the person is dependent on a particular metaphysical doctrine. Though Rawls attempted to configure political liberalism as a fundamentally Kantian project without the metaphysical aspects, Kant argued that people possessed free will and could thus be considered morally responsible. Rawls retains a belief in moral responsibility, yet the grounds on which we consider people morally responsible are not clear. Though moral responsibility is not the central concern of this thesis – it is not responsibility but control that I argue is fundamentally undermined in later chapters – I argue there is a tacit acceptance of the truth of free will in Rawls' theory due to his acceptance of moral responsibility. As Saul Smilansky has argued, though Rawls sets out from a position of hard determinism – goods are not to be distributed according to desert, due to its arbitrariness (Rawls 1971, pp.310-315) – the rest of Rawls' work is "infused with the assumption of free will and moral responsibility" (Smilansky 2003, p.132).

There are two central problems I identify in this chapter. First, due to this tacit acceptance of the truth of free will, Rawls is not remaining neutral in regard to metaphysics. Thus, political liberalism's supposed neutrality between reasonable philosophical doctrines is broken. Second, Rawls assumes people have a degree of control over their thought processes. On Rawls' view, we can freely choose our own ends as rationally autonomous agents, while ensuring these ends remain within the limits of what is just, as we are also fully autonomous agents. If this is the case, then I argue Rawls requires us to hold a large degree of control over our thought processes. It is this view I challenge in subsequent chapters. In order to establish why this is a problem, however, a discussion of moral responsibility and free will in relation to Rawls's theory is necessary. If we are to be the free, responsible agents Rawls supposes us to be, we require a degree of control over our thoughts.

I begin in 5.2 by explaining the dominant positions on free will in metaphysics. In 5.3, I explore how Rawls thinks about freedom in A Theory of Justice. I argue that Rawls' conception of self-respect, which plays an important role throughout Rawls' work, is dependent on a social world that recognises the truth of free will. There is a particular conception of the person that Rawls holds; in 5.4, I argue that this is still the case with *Political Liberalism*. Despite Rawls' aim to configure political liberalism without the need for metaphysics, in 5.5, I argue that there are nevertheless metaphysical considerations underlying political liberalism. I argue that Rawls' thought is dependent on the human capacity for free choice, and that it thus presupposes the truth of free will. Though Rawls attempted to formulate Kantian politics without the metaphysical component, Rawls retains certain metaphysical assumptions. This is particularly problematic in regard to Rawls' conception of rational autonomy; if we cannot formulate a conception of the good in the way Rawls imagines, aspects of Rawlsian theory should be revised. As it is necessary for citizens to become fully autonomous, we must consider how it is we expect people to do so.

5.2 Metaphysical Perspectives on Free Will

There are three perspectives on free will that dominate discussion on the metaphysics of free will: hard determinism, compatibilism, and libertarianism. I now explore these three perspectives, and the various ways in which they have been formulated. The aim of this section is to provide an explanation of some key ideas within the subject, and an understanding of the terminology.

Before explaining hard determinism, it is necessary to understand what is meant by determinism. The thesis of determinism states that from one moment in time to another there is only one possible eventuality (Fischer et al 2007, p.2). Someone accepting determinism endorses the view that in the universe, at any given moment in time and in any situation, one event will necessarily lead to another event, and this could never have been otherwise. Everything is subject to the laws of physical causation; all of the events that occur throughout time and space are necessitated by the events that preceded them. This is true at all levels of matter, from the sub-atomic level upwards (Dennett 2003, pp.97-101).

The first philosopher to notice the antinomy between determinism and free will was Epicurus (Weatherford 1991, p.19; Epicurus 2004), who modified his account of determinism to allow for human free will. Others think that determinism negates the possibility of free will completely. *Hard determinism* is the view that determinism is true, and as a result, we do not have free will. The laws of nature, and of physical causation, determine how any event unfolds. Human actions are no different. We had no control over past events, nor do we have any control over the laws of nature (Van Inwagen and Griffiths 1985). Therefore, we have no choice in regard to the actions we make. The hard determinist argues we could never have acted other than we did. As our inner mental states are the result of laws over which we have no control, we cannot be the original creators of our own ends, making our inner motivations irrelevant in regard to the question of free will.

This can be illustrated through the example of Laplace's Demon (Weatherford 1991, pp.52-60). While Isaac Newton's understanding of mechanics is generally thought to be deterministic, in his theory of gravity, Newton left space for the existence of God, who was thought to occasionally intervene in the workings of the universe. Roy Weatherford writes that it was Pierre Simon de Laplace who took Newtonian mechanics to a deterministic conclusion (Weatherford 1991, pp.52-60; Laplace 1814), as witnessed in the idea of his Demon. Laplace's Demon holds a complete understanding of the universe. From knowledge of the position of all the particles in the universe at any one moment, Laplace's Demon can predict with certainty both the particles' past and future. Thus, from being given a complete picture of an individual's place in the universe at a moment in time, the Demon could tell the individual her future. The individual cannot be viewed as holding any meaningful control over her life, given that the Demon knows with certainty her entire life story, there being nothing she can do to alter this story. This view of the universe holds no place for the concept of free will.

Compatabilism accepts that determinism is true, or at least could be true, but the compatibilist argues that determinism does not negate free will. Ted Honderich argues that Hobbes was the first philosopher to posit a theory of compatibilism (Honderich 2002, p.105-108). Arguing that determinism negated free will, Bishop Bramhall provoked Hobbes into formulating the compatibilist defence (Hobbes & Bramhall 1999). For Hobbes, a person was to be considered free providing that nothing was externally restricting her actions. David Hume followed Hobbes in presenting the case for compatibilism. Accepting the probable truth of determinism, Hume argued that the supposed problem of free will was merely a problem of language and could be rectified with "intelligible definitions" (Hume 1748, p.59). Going on to define liberty as "a power of acting or not acting, according to the determinations of the will" (Hume 1748, p.125), Hume argues that everybody holds the capacity to choose whether or not to act, apart from a prisoner in chains. Early modern compatibilists, such as Hobbes and Hume, thought that determinism did not negate free will as to have free will was only a matter of the will being determined in a particular way. Whether or not determinism is true, providing we are not externally restrained, we have the capacity to choose. Therefore, for Hobbes and Hume, free will is compatible with determinism.

Rather than asking whether the agent could have done otherwise from a metaphysical perspective, we need only enquire whether anything was preventing the agent from acting otherwise: was the agent physically restrained, threatened, or were her mental capacities impaired. However, ascertaining whether we could have done otherwise does not necessarily determine whether or not we have free will for the compatibilist. Instead, some compatibilists instead consider whether an act was in accord with our inner motivations. Illustrating this argument, John Martin Fischer devises a thought experiment based on those of Harry Frankfurt (Fischer 2012, pp.33-35; Frankfurt, 1969). Fischer asks us to imagine a person voting in an election in the United States. The woman intends to vote for the Democratic candidate. However, a neuroscientist has placed an implant in this person’s brain. If they decide to vote for the Republican candidate instead, the implant will intervene to ensure the person votes for the Democrat. If the person stays true to the original intention of voting for the Democrat, the implant will do nothing. The woman does decide to vote for the Democrat; thus, the implant does nothing. While the woman remains unaware of the implant, she could never have done otherwise than she did. Frankfurt's original response to this dilemma was that it did not, though Fischer modifies his own response. According to Frankfurt, rather than assess whether a person could have done otherwise than she actually did, we should instead look to inner motivational states, which are the most important factor in determining whether a person acted according to her free will (Frankfurt 1969, pp.837-839). An action that is in accord with a person’s inner motivational state is an act of free will, whether or not that person could have done otherwise. Therefore, Frankfurt argues that free will is not dependent on the falsity of determinism. Fischer, on the other hand, states that this is not sufficient for free will, effectively agreeing with the hard determinist, though he argues that it does suffice for the ascription of moral responsibility. Fischer labels this position semi-compatibilism: free will is incompatible with determinism though moral responsibility is not.

Kant can also be interpreted as a proponent of compatibilism, though Kant's compatibilism is rather different from the aforementioned versions of compatibilism. Compatibilism, as formulated by Hume and Hobbes, was for Kant a "wretched subterfuge" (Kant 1788, p.78). In the same vein as the hard determinist, Kant argues that the thief who could not have avoided being a thief due to the laws of determinism could not be viewed as having freely chosen to be a thief. Hume and Hobbes, by way of their "wretched subterfuge", were avoiding the implications of determinism's truth. The notion of transcendental freedom enables Kant to argue for a formulation of free will that is compatible with determinism, yet not dependent on the reconciliation of freedom and necessity. For Kant, it was true that, empirically speaking, all things are determined causally (Kant 1781, p.169). This empirical evidence, however, is only an appearance. The way in which we witness the workings of causal determinism belongs to the phenomenal realm. In the phenomenal realm, we see the thing as it appears. It is in the noumenal realm that we find the thing as it is. The laws of causal determinism are not applicable to the noumenal realm. This distinction allows for transcendental freedom. Our transcendental selves belonging to the noumenal realm are not subject to causal laws and are therefore free (Kant 1781, pp.405-408). Kant offers a unique conception of free will that is compatibilist insofar as it reconciles freedom with determinism, though transcendental freedom is far removed from a Humean understanding of free will.

Libertarianism is the view that we do have free will. Like the hard determinist, the libertarian views free will as being incompatible with determinism. To have free will, not only must we be able to act otherwise, we must be the creator of our own ends. Determinism would mean that neither of these propositions is true, and, as we have free will, determinism must be false. Though he is not the only contemporary proponent of metaphysical libertarianism, Robert Kane has formulated what has become the dominant theory of libertarianism. Kane argues that to consider an agent as having acted from her free will, she must hold “Ultimate Responsibility” for this action (Kane 1996, pp.72-75).

To hold ‘Responsibility’ means to have acted voluntarily, and to be involved in the causal determination of an event. Without the agent’s role, the event would not have occurred. However, this level of responsibility does not imply “Ultimate Responsibility”. For the condition of “Ultimate Responsibility” to be met, an agent must be the original creator of her own ends, and her ends must determine the actions she chooses to make. Thus, if an agent carried out an action due to her Protestant faith, she must also be responsible for deciding to adhere to Protestantism. “Ultimate Responsibility” requires that if an agent is responsible for X, but Y was also necessary for the occurrence of X, then the agent must also be responsible for Y (Kane 1996, pp.72-75.). Thus, if a particular mental state was required in order for a person to act, for this act to be an act of free will, the person must also be responsible for freely choosing that mental state. We form these mental states through what Kane calls “self-forming actions” (Kane 1996, pp.75-76). When we are faced with a moral dilemma, our decision in regard to the dilemma becomes a “self-forming action”: through making this decision, we have decided upon a crucial aspect of our own character. With these conditions met, the agent then holds “Ultimate Responsibility”. Kane follows Aristotle in arguing that if our actions are judged to be free, it must be because they stem from our characters, which we have formed through our own free choices (Kane, 2011, p.383). Our characters provide the roots from which free will stems; thus, the decisions that brought about our characters ground how Kane formulates responsibility. Kane recognised that the theory of “Ultimate Responsibility” could lead to an infinite regress, but argues that it need not entail events prior to a person’s birth nor regress further than “self-forming actions” (Kane 1996, pp.72-76).

If determinism is true in all cases, then the agent could never hold “Ultimate Responsibility”. In any scenario, if there were always a Y that caused X, and this Y was beyond the control of the agent, then the agent could not be considered to have free will. In a deterministic universe, there would always be a cause antecedent to the agent’s own intentions and actions, negating the sense in which she could be considered the creator of her ends. We could not hold responsibility for the creation of our ends if Laplace’s Demon knew our ends even before our own births, and we could choose no other ends. Because determinism is not true, according to the libertarian, there is the possibility of free will. Thus, the libertarian argues there are ways in which a person can be the creator of her own ends. As we can freely create our own ends, we can freely choose our actions based upon these ends. Kane’s libertarianism is demanding of free will; not only must we be able to act otherwise, we must be able to freely choose our own ends, and make free actions based upon these ends.

Though determinism is not true according to the libertarian, as Kane notes (Kane 1996, pp.106-107), the flipside of arguing for an indeterminist account of free will is that if all of our actions are indeterminate, this looks as though they are subject to chance, which hardly suggests we have control over these actions. This was why Hume thought that not only did determinism not negate free will, it was a necessary condition for free will (Hume 1748, pp.58-59). In order for us to have control over events, there must be uniformity in the predictability of outcomes. Free will, for Hume, consisted in our will cohering with deterministic laws in the right way. For the plausibility of libertarianism, then, the inner workings of the brain can neither be wholly deterministic or indeterministic, insofar as complete indeterminism would be total randomness. Drawing on evidence from quantum mechanics and neuroscience (Kane 1996, p.9), Kane argues science has not proved determinism to be true, and it is plausible that the right level of indeterminacy exists in the human brain for the truth of libertarian free will. Our actions cohering with our will, which has been formed through our "self-forming actions", reflects this sufficient level of indeterminacy, negating the sense in which our actions can be said to be a result of chance.

Derk Pereboom, on the other hand, takes the view that free will is neither compatible with determinism nor indeterminism, labelling this view *hard incompatibilism*. Following Hume, Pereboom argues that indeterminism suggests that all events are due to chance. Using Kane's example of a businesswoman considering whether to stop to help the victim of an assault or attend a meeting on time, Pereboom argues that the decision made is dependent on chance. If we imagine that there is equal motivational force for either action, the action eventually decided on will be a chance result (Pereboom 2007, pp.101-103). This level of chance does not suggest free will played an important role in the action. Elsewhere, Pereboom agrees with the hard determinist and the libertarian that free will is incompatible with determinism. Pereboom agrees with David Widerker's rebuttal of compatibilists such as Frankfurt (Widerker 2000; Pereboom 2007, pp.87-92). While Frankfurt thought that a person with no alternative possibilities could still be held responsible for the action she committed, Widerker asks the proponents of this view to tell us what else she could have done. Pereboom argues that there is no suitable response to Widerker here. The agent could never have done otherwise, and the fact of coherence between her will and her actions does not provide sufficient grounds for her will to be free. Thus, neither determinism nor indeterminism can allow for the existence of free will.

This completes a brief description of each metaphysical perspective. Though there are other theories of free will, such as Saul Smilanksy's illusionism and Manuel Vargas' revisionism (Smilansky 2000; Vargas 2004), these alternative theories are broadly compatibilist – arguing for the compatibility of determinism and free will – but focus instead on the precise way in which free will is formulated.

To conclude, libertarians and compatibilists agree we have free will; however, they disagree on the precise ways in which we have free will; while the compatibilist views free will as independent of the issue of determinism, libertarians argue that in order for free will to be true, determinism must be false. Hard determinists agree with the premise of libertarianism – determinism negates free will – though the hard determinist inverts the argument, and states that because determinism is true, we cannot have free will.

5.3 Freedom in A Theory of Justice

In this section, I explain Rawls' thought on freedom in *A Theory of Justice*. Throughout his work, Rawls rarely approached the subject of free will, and when the subject comes into view, it is generally sidestepped. The only sustained discussions of free will in Rawls' oeuvre are found in his analyses of the works of others, such as Leibniz, Kant, and Hegel in *Lectures on the History of Moral Philosophy*, and Rousseau in *Lectures on the History of Political Philosophy* (Rawls 2000; Rawls 2007).[1] Though Rawls is concerned with freedom in *A Theory of Justice*, he considers political freedom to be distinct from metaphysical freedom. I argue that Rawls explicitly endorses a thin form of compatibilism in his definition of freedom. Rawls, following Hobbes and Hume, argues that a person can be considered to be free providing there are no external restraints. However, within the original position, Rawls tacitly endorses Kantian compatibilism, as shown in the Kantian interpretation of justice as fairness.

Defining freedom, Rawls attempts to avoid becoming embroiled in the debate between negative and positive liberty, and instead explains freedom through three items, allowing for a formulation of freedom that encompasses both positive and negative liberty (Rawls 1971, pp.201-202). Freedom concerns the agent who is free, the restrictions she is free

[1] G.A. Cohen notes that, in private conversation between them, Rawls expressed doubt whether the moral worth of people would hold if all actions were causally determined (Cohen 2008, p.14). Cohen was pleased that Rawls appeared to share his scepticism regarding the compatibilist position on free will.

from, and the acts that she is free to do. Freedom is lexically prior to equality throughout Rawls' work, with the first principle of justice taking precedence over the second. The reason Rawls provides for freedom's priority is the importance generally ascribed to freedom in modern societies. Rawls writes that it is usually accepted that there is competition between the values of freedom and equality, and that, in general, preference is given to freedom (Rawls 1971, p.28). Freedom, as Rawls is attempting to avoid becoming tied to a particular philosophical definition of freedom, is understood as the attribution of basic rights and liberties. People would not be willing to cede their liberties in favour of a more egalitarian society. Basic liberties protect the interests of citizens; under a scheme of basic liberties, a citizen can be sure that they have the right to practise her religion, or to pursue a plan of life in accordance with her own personal values (Rawls 1971, pp.205-207). Thus, the basic liberties are lexically prior to the ensured social and economic equalities.

Both the citizen of the well-ordered society and the parties in the original position are considered to be free. The citizen is free as the first principle of justice safeguards her liberty. Parties in the original position are intended to be free, constrained in their deliberations only by the conditions imposed in the original position (Rawls 2005, pp.74-75). Rawls takes an implicitly compatibilist approach here; as the choices are made by people who exist in equal relations to one another, and there are no external obstacles to their making choices, the choices can be considered to be free. No commitment is made, however, to the deeper metaphysical aspects of compatibilism. Rawls does not, for instance, suggest that a person's inner motivations must be aligned with her actions in order for her to be considered free. The stance assumed here recalls the compatibilism of Hobbes, rather than compatibilism with its deeper metaphysical commitment to the noumenal self in Kant, or Frankfurt's argument for coherence between psychological motivations and actions. Nevertheless, through the original position, Rawls retains elements of Kantian compatibilism, as I explain later.

In *A Theory of Justice*, Rawls includes a Kantian interpretation of justice as fairness (Rawls 1971, pp.251-257). Rawls claims that the principles of justice are comparable to Kant's categorical imperative. The principles have universal applicability in the same sense as does the categorical imperative. When we come to act from the principles, we act from a universal moral law that is applicable to all, regardless of our own interests. Here, Rawls' view of duty and obligation in relation to the principles of justice begins to resemble Kantian autonomy, as Rawls himself notes. However, following Sidgwick (Sidgwick 1888), Rawls does not accept the conception of the self in the distinction Kant draws between the noumenal and the phenomenal.

Kant thought that the self was only realised when the person acted from the moral laws (Kant 1797). When a person acts against the moral laws, she acts according to the phenomenal laws of nature, as the desires that led her to disobey the moral laws belong to the phenomenal realm. A person acting in accord with moral laws recognised through the use of reason would not act egoistically. The noumenal self, a self that is detached from the phenomenal realm and able to comprehend the moral laws, remains unrealised when a person acts against the moral laws. According to Sidgwick's reading of Kant, the saint and the scoundrel both freely choose their characters in the noumenal realm (Sidgwick 1888, pp.409-410), and these characters are then subject to the same physical laws in the phenomenal realm. Sidgwick argues that Kant does not explain why the scoundrel does not express his freely chosen self in the same way as does the saint, given that their characters are both the result of the same free choices. Why, then, can we not consider the scoundrel as free?

Rawls suggests that the answer to this problem is found in the original position (Rawls 1971, pp.251-257.). Parties in the original position look upon citizens in the same way that the noumenal self views the world. As we can come to understand the importance of self-respect through the original position, we can realise that failure to act in accordance with the principles of justice would detract from our self-respect, and lead to a feeling of shame. Due to Rawls' belief that all rational persons would endorse the principles of justice in the original position, this sense of shame has universal applicability. No one would agree to the scoundrel's terms in the original position, and so the person who chooses to act as a scoundrel is not doing so based on the free choices made in the original position.

An answer is thus given to the problem identified by Sidgwick. Through the original position, Rawls strips Kantian morality of its metaphysical counterpart (Rawls 1971, pp.251-257). The distinction between the noumenal and phenomenal selves is then abandoned within Rawlsian thought, though the original position provides an effective substitute. Rather than thinking about the relationship between the self and the world from the perspective of Kantian metaphysics, Rawls allows us to consider the relationship through the original position, devised as a thought experiment rather than a metaphysical claim.

However, at the heart of the distinction between the phenomenal and the noumenal is the problem of free will, which Rawls does not consider through the substitute for this distinction, the original position. Kant thought that the noumenal self salvaged the freedom of the person's will, as without it, the phenomenal self would be determined only by the laws of physics (Kant 1788, pp.42-43). Unless it is assumed that the original position offers us the opportunity to transcend the restrictions placed on us by the laws of physics, much in the same way as does the noumenal self, Rawls must accept that, if determinism is true, the self will be physically determined. Thus, our sense of justice and conception of the good will be arrived at through deterministic physical laws, rather than our own capacity to reason creatively, independently of the laws of causation.

To establish why is it necessary the original position salvages free will, we need to understand why freedom is important for Rawls. As previously mentioned, freedom is prioritised within justice as fairness because of people's general preference to be free. Rawls also writes that a regime not upholding basic liberties could not be considered a constitutional democracy (Rawls 1971, pp.197-198). Though Rawls hints at some of the psychological and historical factors influencing our preference for freedom, he does not attempt to uncover why people prefer to be free over other goods. Freedom's importance in *A Theory of Justice* lies in ensuring our access to the primary goods, particularly in relation to the realisation of self-respect. When we construct a rational plan of life and then witness our successes in light of this plan, our self-respect is enhanced (Rawls 1971, pp.440-446). Our failure to construct a rational plan of life leads us to experience little joy in life, and we are left without a sense of value or purpose in our life. Furthermore, we take pride in our plans when we receive praise, and we also find satisfaction in the successful plans of others. Failing to live our life according to a rational plan leads to a feeling of moral shame; we experience this not only because of our own reduced satisfaction with life, but also because of the negative way in which others will regard us. Actions committed that contradict agreements made in the original position will be censured by the rest of society. It is therefore in our interest to live according to a rational plan of life that is agreeable to the rest of society, and enhance our feeling of self-respect.

The sense of worth we experience due to our life plans' fruition reflects pride in our capacity to choose, and our satisfaction that we exercised this capacity virtuously. We essentially chose to be good rather than bad people, though the option was open for us to have chosen otherwise. If we were told that we live in a deterministic universe, and that Laplace's Demon could have foretold our characters, our plans of life, and all of our actions before the time of our own births, we would have no reason to feel this sense of worth, as it could never have been otherwise. In such a universe, whether we are good or bad people is not due primarily to the free choices we have made; rather, it is due to the course of the universe, over which we have no control. The fact that we act in accordance with rational plans of life and experience success, while others act against the agreements made in the original position, would not be reflective of our own agency and our ability to take responsibility for our actions. Instead, all of our characters, plans, and actions would be predetermined, and we should feel neither pride nor shame in our successes and failures. Though we may still experience happiness when our plans are successful, we would recognise that, ultimately, this plan could never *not* have been successful. Likewise, we could not castigate others for their moral failings, and though they may still feel shame due to these failings, it would also be recognised that they could not have done otherwise. Determinism's truth would not entail a world devoid of emotion, in which no pride or shame could be felt. However, it would be recognised that these emotions were only natural reactions to events, and lacked the deeper moral justification that hinges on the attribution of responsibility.

In other respects, however, Rawls recognises the limits of free choice. Rawls rejects the notion that goods should be distributed according to moral desert, where those who contribute more receive more in return (Rawls 1971, pp.310-315). This is partly because the talents involved in making the contribution are developed according to the arbitrary factor of birth. Some people will be born with a talent for singing, while others will not, and it would therefore not be just to award the person possessing the talent, as the person did not choose to possess this talent. The idea of moral desert would then be rejected in the original position, as, according to Rawls, we would not accept distribution according to such arbitrary factors (Rawls 1971, p.310). Here, Rawls' thought is largely compatible with hard determinism; economic and social structures should be arranged only so as to encourage certain choices to ensure the efficiency of these arrangements, not to reward those who contribute more due to their talents (Rawls 1971, p.315). However, as I later argue, certain actions made as a result of free choices can affect the economic and social structure of a society, and through so doing, violate the second principle of justice.

Though Rawls, in *A Theory of Justice*, quite explicitly accepts the thin compatibilism of Hobbes, by arguing that we are free if nothing is externally restricting us, in the construction of the original position, Rawls tacitly accepts Kantian compatibilism. If we are to accept the senses of self-respect and shame that come with either following or failing to follow a rational plan of life, we must accept that people can freely choose, and can do otherwise. The original position allows us an exterior vantage point, akin to Kant's noumenal realm, from which we can imagine people as being free. As these plans are the result of choices people have freely made, they are justified in feeling pride in their success.

5.4 Political Liberalism and the Metaphysics of Free Will

Political liberalism's neutrality between reasonable metaphysical doctrines should lead it to be independent of any theories of free will. The politically liberal state should not be committed to a particular metaphysical position on free will, and the citizens of this state should be free to endorse whatever doctrine they find convincing. In this section, I explain Rawls' own position on free will, elaborating on why the metaphysics of free will would be excluded from the scheme of political liberalism. Following this, I explore certain claims made within Political Liberalism that bear on the metaphysics of free will. Despite its aim, political liberalism does not remove the compatibilist position found in *A Theory of Justice*.

Before publishing *Political Liberalism*, Rawls presented his reasons for avoiding metaphysical discussions in 'Justice as Fairness: Political not Metaphysical' (Rawls 1985). Many metaphysical and moral problems are viewed by Rawls as being intractable (Rawls 1985, pp.226-227). The social contract existing at the heart of a democratic regime should not be devised so as to be dependent on a particular solution to such an intractable problem, as the problem could not be solved through that regime's conception of justice without sacrificing the liberty of citizens to determine their own beliefs, further dividing society in the process. In formulating how a well-ordered society could be realised through a conception of justice agreeable to the majority of citizens, the political theorist should not draw on metaphysical arguments. This is not to say that these concerns are unimportant; Rawls stresses that, in fact, they are “too important” to be approached through political deliberation (Rawls 1985, p.230). However, while Rawls avoids attempting to settle metaphysical disputes, there are metaphysical assumptions within Rawls’ work.

In *Political Liberalism*, Rawls further distances himself from metaphysical commitments. The conception of justice at the heart of the politically liberal state must not be dependent on a particular metaphysical doctrine. Decisions made in the original position should, then, be neutral between the metaphysical positions on free will. Hard determinism, compatibilism, and libertarianism would all be viewed as reasonable doctrines, as the objective of any one position is not to remove the other positions from society, or to use the power of the state to suppress other metaphysical doctrines. A citizen would be free to endorse the metaphysical position she found most convincing, though no one metaphysical position should be used to legitimise the conception of justice.

While Rawls is not particularly concerned with the metaphysics of free will in his explanation of public reason – as he is mainly considering other philosophical, religious, and moral doctrines (Rawls 2005, pp.212-254), aspects of thought over which, historically, people have been divided – it is nevertheless safe to assume metaphysical perspectives on free will would be excluded from public reason. As Rawls writes, there is no reason why decisions on constitutional essentials ought to be based on reasons drawn from a particular comprehensive doctrine (Rawls 2005, pp.25-26). Therefore, the state’s constitution should not favour one metaphysical perspective on free will over others. Proponents of particular metaphysical views would also be prohibited from using the apparatus of the state to promote their views or suppress other views. Though it may seem unlikely that a metaphysical libertarian would wish to use the power of the state to suppress hard determinism and compatibilism, these perspectives on free will are also contained within other religious and philosophical doctrines (Kane 1996, pp.7-8). Historically, these doctrines have been hostile toward one another. It is plausible to imagine that if the proponents of one of these doctrines controlled the power of the state, they would use it to the advantage of the metaphysical perspective belonging to their religious or philosophical tradition. Free will is not a subject beyond political contention.

Despite the fact that these metaphysical doctrines would be omitted from the public reasons used to support constitutional essentials, Rawls retains certain concepts from *A*

Theory of Justice. I now turn to how Rawls defends the conception of the person he devises in *Political Liberalism*.

In response to his communitarian critics, Rawls stresses that the original position is to be viewed as a thought experiment. It is therefore not expressing a certain perspective on the nature of the self. Rawls writes that when:

> We simulate being in the original position, our reasoning no more commits us to a particular metaphysical doctrine about the nature of the self than our acting in a play, say of Macbeth or Lady Macbeth, commits us to thinking that we really are a king or queen engaged in a desperate struggle for political power (Rawls 2005, p.27).

However, Rawls still has a particular conception of the self within the original position. Self-respect remains the most important primary good. Our motivation to endorse the principles of justice stems from our desire to advance our self-respect and the realisation that the principles of justice will realise this desire (Rawls 2005, p.318). The motivation to become fully autonomous is primarily built upon our self-respect's advancement. Without self-respect, we will not be motivated to endorse nor act from the principles of justice, and our moral powers will remain undeveloped. Hence, the scheme of basic liberties is necessary to ensure the conditions required for the development of our self-respect are met.

Nothing here has changed from what was posited in *A Theory of Justice*. Self-respect remains necessary for the construction of a society based on terms of fair cooperation, in which fully autonomous citizens will live cooperative lives, ensuring the realisation of their individual talents and conceptions of the good. Without self-respect, none of this is possible.

Rawls anticipated the Hegelian critique of his project. Hegel, as an idealist, criticised the contractarian doctrines of Hobbes and Locke for two reasons, of which the second is pertinent here (Rawls 2005, pp.285-288; Hegel 1821). Humans, in the thought of Hobbes and Locke, were thought to have innate characteristics, existing independently of society. For Hegel, humans are social animals and the characteristics we hold are a result of the societies in which we exist. We do not, according to Hegel, hold characteristics that are independent of societal influences. Rawls countered this argument by stating that political liberalism was perfectly capable of accounting for the social nature of individuals; he writes that justice as fairness is “a moral conception that provides an appropriate place for social values without sacrificing the freedom and integrity of the person” (Rawls 2005, p.286). We can hold whatever social and communal values we like, providing that these values are reasonable, and do not override our motivation to endorse the principles of justice.

Unacknowledged in this defence, however, are the characteristics Rawls continues to attribute to humans, without which, political liberalism would be unworkable. This perspective on the individual, in which we are viewed as attaining self-respect through acting from the principles of justice, is dependent on an individual accustomed to certain social and cultural norms. A person in this society attains her self-respect from realising her rational plan of life, and the realisation that her plan acts toward advancing the aims of a fairly cooperating society. Nowheresville, Joel Feinberg’s thought experiment, further clarifies the importance of self-respect through the imagining of its absence (Feinberg & Narveson 1970). This imaginary location, Nowheresville, is a place in which the inhabitants do not have rights. They therefore are unable to make claims based upon rights. Feinberg claims that in a world in which people could not make claims based upon their rights, people would be unable to realise their self-respect (Feinberg & Narveson 1970, p.257). It is through the ability to make these claims that we come to realise our inherent worth as a human being. Without this ability, we would see no value in self-respect. The members of such a society would, then, not find the attainment of self-respect a sufficient motivational force to compel them to commit to the social contract. Hence, for Feinberg, rights are necessary preconditions for self-respect; it is only through our possession of rights that we are able to realise our self-respect.

For the realisation of autonomy in the politically liberal state, self-respect must be an inherent part of the human condition. Instead, as Feinberg shows, self-respect is constructed through the attribution of rights. A person incapable of realising her own self-respect, or even realising the importance of self-respect, would not be considered a fully autonomous citizen within political liberalism. Furthermore, as previously mentioned, self-respect is a value representative of a world in which we recognise the importance of free choice. Thus, there is a particular conception of the person situated within political liberalism that is dependent on certain social arrangements existing prior to the citizen.

With rights being prior to self-respect, we cannot assume that individuals consider themselves as possessing a sense of self-respect from the perspective of the original position, nor imagine that they will necessarily enhance their self-respect through the development of their moral powers. Parties in the original position are denied knowledge of the citizens' historical circumstances. They could, then, not be certain of the psychological attributes of citizens. If we are to imagine the veil of ignorance as being effective, then we must not make assumptions about psychological states that are historically contingent.

Instead, Rawls assumes that psychological states are independent of history. Whatever our historical circumstances, we are capable of developing the sense of morality Rawls imagines we hold. Thus, Rawls must consider the two moral powers and the sense of self-respect as innate characteristics. From these powers, we have the ability to make free choices for which we can be held morally responsible. It is this that allows for us to be rationally autonomous – capable of realising our two moral powers (Rawls 2005, 72-77). The notion of rational autonomy is, then, undermined if the truth of free will is rejected, as through its rejection, we also reject the idea that people hold the ability to make free choices for which they are to be responsible. Full autonomy – our capacity to act from the principles of justice (Rawls 2005, p.77) – is less vulnerable to this critique.

In sum, though in *Political Liberalism* Rawls attempts to show that justice as fairness is not dependent on a particular metaphysical conception of the self, in the importance Rawls assigns to the role of self-respect, we find a certain perception of the person. This perception is based on the individual as she exists in a particular social group, in which certain values are expressed as social norms. Without the acceptance of free will and moral responsibility in this social world, the motivational force of self-respect would be neutered. While full autonomy is not necessarily undermined by this argument, rational autonomy is threatened in a universe without free will.

5.5 Autonomy and Free Will

Rawls is, then, dependent on a particular conception of human nature. The person can choose her own good, and develops her sense of self-respect when she does so. If citizens are to become fully autonomous in the way Rawls imagines, we must choose to do so through our capacity for rational autonomy. In the way in which Rawls imagines us making such choices, I argue that Rawls is dependent on the truth of free will. Without the truth of free will and moral responsibility, we would need to revise how we think about rational autonomy, and how we imagine citizens becoming fully autonomous. Though I focus here on the problem of free will and moral responsibility, the central problem I aim to uncover here is the demanding conception of psychology Rawls holds. For us to attain rational and full autonomy in the way proposed by Rawls, we require a level of control over our thought processes.

In a defence of Rawls' position, Veljko Dubljevic (2013) has argued against the conflation of metaphysical free will with political autonomy. Dubljevic's definition is, however, quite different from Rawls'. In his definition of political autonomy, Dubljevic essentially commits to the compatibilist position. This reveals the difficulty of remaining metaphysically neutral when defining autonomy. Dubljevic writes:

> An agent acts autonomously when she/he (a) endorses decisions and acts in accord with internal motivational states, (b) shows commitment to them in the absence of undue coercion and compulsion, and (c) could as a reasonable and rational person continue to do so after a period of informed critical reflection (Dubljevic 2013, p.46).

Nothing here contradicts or distinguishes this position from Humean compatibilism. When Hume says that anyone is free to choose to rest or move asides from the prisoner in chains, the argument is premised on the causal relationship between the will and act. In stating that an act is autonomous if there is a relationship between the "internal motivational state" and the act, it is an autonomous act, Dubljevic commits to Humean compatibilism. The two additional conditions – lack of coercion and compulsion, and endorsed on reflection – do not further distinguish Dubljevic's position from Hume's. An act that was the result of a violent threat or delirium that the person regretted in reflection would presumably not be thought of as an act of free will, on Hume's view. The essential part of the argument, for Hume, is the connection between the will and the act, with the act revealing the contents of the will. Where this connection does not hold, we could not view the act as being made of free will. In much the same way, this act would not be considered autonomous on Dubljevic's view. It is not clear that there are cases that would be considered free by Hume, but not autonomous by Dubljevic, or vice versa. Furthermore, although Dubljevic draws on the Rawlsian conception of autonomy with its lack of a metaphysical component, Dubljevic does not distinguish between rational and full autonomy. Neither are adequately captured by the above definition. A person's acts could cohere with her will while she neither acted to realise her moral powers (as in rational autonomy), nor from the principles of justice (as in full autonomy). Rawls' two definitions of autonomy are more demanding than Dubljevic recognises: we need to be capable of doing more than ensuring our acts cohere with our will to be considered autonomous on Rawls' view.

It is Rawls' notion of rational autonomy that is problematic when assessed against the metaphysics of free will. If we are to be considered capable of formulating a conception of the good, taking responsibility for the ends arising from this conception (Rawls 2005, pp.72-77), we need to be considered as being able to freely choose these ends. As I explain below, if we are assumed not to possess free will, Rawls' conception of rational autonomy cannot hold. On the other hand, full autonomy does not face such problems. Whether or not we possess free will, we may be able to act justly. As Rawls recognises, the attainment of full autonomy is an epistemic problem (Rawls 2005, p.78),[2] not a metaphysical one. Providing we know what is just and, in turn, what is expected of us as citizens, then we can be fully autonomous citizens. The capacity for making free choices is less problematic for full autonomy than it is for rational autonomy.

Against Dubljevic, William Simkulet argues that political liberalism presupposes metaphysical libertarianism, as the current liberal democratic state as it exists is dependent on metaphysical libertarianism being true. To illustrate this, Simkulet asks us to imagine that, in the original position, we are to suppose that hard determinism is true, and that humans have no free will (Simkulet 2013, p.72). Under such conditions, parties would devise laws that would be radically different from the laws of most liberal democracies. Whether we are deemed to be wholly responsible for our actions matters under most legal systems. If we are deemed never to be entirely responsible, Simkulet argues that laws should be devised so as to reflect this. That Rawls does not come to this conclusion represents, for Simkulet, the fact that political liberalism is dependent on metaphysical libertarianism. For Rawlsian justice to be realised, citizens must be deemed capable of making free choices and taking responsibility for these choices. However, though Simkulet posits this as a commitment to libertarianism, it is better characterised as a dependency on the truth of free will, however conceptualised. Simkulet's argument does

[2] It is only with the full publicity condition being met that we could expect a person to become fully autonomous (Rawls 2005, p.78).

not account for compatibilism; it is plausible to imagine determinism being thought of as true within the original position, but the parties agreeing to accept the truth, or plausibility, of compatibilism.

As Simkulet states, for us to be deemed responsible for our rational plans of life as Rawls hopes, we must be thought to be free. In a social world in which hard determinism was accepted as true, we would not attain this level of freedom on which we could be thought responsible. Accepting the truth of hard determinism would not radically transform human life, according to Ted Honderich (1973, p.213). Our views on individuality and responsibility, however, would have to be revised. Honderich writes that way in which the majority of people think about individuality and responsibility is premised on a belief in free will (Honderich 1973, pp.208-209). With the truth of hard determinism, this would need to be amended to account for the fact that we would now know that a person could never have done otherwise. As the *telos* of political liberalism is not the realisation of individuality in the sense that it is in J.S. Mill's *On Liberty* (1859), re-evaluating how we think about individuality would not pose a concern for political liberalism. However, responsibility, while not the ultimate end of political liberalism, does play a significant role in Rawls' account of autonomy. If we cannot be deemed responsible for our plans of life, we cannot attain rational autonomy.

How would responsibility, then, need to be reframed according to hard determinism? For Derk Pereboom (arguing from the hard incompatibilist position, which, however, overlaps with the hard determinist view), wrongdoing should be considered in much the same way as a natural disaster; while it may be regrettable, the person guilty of wrongdoing could not have done otherwise, and so we should not seek retribution (Pereboom 2006, p.154). A more pragmatic approach would need to be taken towards human wrongdoing, moral failures, and poor decision-making. Violent criminals would still need to be imprisoned to protect the rest of society, yet this would be a matter of security rather than the punishing of a person. We would also need to recognise that poor decisions were not the fault of the individual person. Hence, the person who gambles away her life-savings cannot be deemed responsible for doing so, nor can the person who joins a racist political movement or endorses a violent, extremist religion. Rather than put the onus on the individual person to not perform such actions, it should instead be society's duty to reduce the likelihood of such actions being committed.[3]

That Rawls does not draw such conclusions suggests, as Simkulet argues, that he is reliant on the truth of free will. As stated above, this is not necessarily a problem for Rawls' conception of full autonomy. Our ability to act from just principles may instead be enhanced by a lack of free will. If a person had been indoctrinated to act from certain principles, the chances of her acting against these principles may be reduced if she does not have free will. In a deterministic universe, it seems safe to assume that a person will be less likely to act against the doctrine under which she was raised. Thus, the problem here is not for full autonomy. Rather, it is rational autonomy that is undermined if we do not have free will. There is an additional problem, however, in regard to how Rawls posits that we become fully autonomous. We come to accept the principles of justice through realising the principles promote our conception of the good (Rawls 2005, pp.144-150). To put this in terms of autonomy, we come to be fully autonomous partly through our rational autonomy. Lack of free will holds implications for how we imagine a person formulates her conception of the good, and whether we can consider her responsible for this conception. This is problematic as Rawls leaves it to the individual to determine her own good.

As Joseph Raz explains, Rawls is not entirely indeterministic in how he imagines we come to hold our conceptions of the good, or the way of life we derive from this conception (Raz 1986, p.131). Instead, Rawls recognises that the sense of morality we hold, and the way of life we choose, will be largely socially determined. We do not need to be capable of making radical choices in which we realise these aspects of ourselves independently of our social world. Nevertheless, we still need to be capable of choosing reasonably and rationally from the choices available to us within this world. A person incapable of doing this could not be considered rationally autonomous.

If we expect a person to become fully autonomous, yet respect her ability to determine her own good, we place a greater demand on her than Rawls recognises. In a society in which freedom of conscience is respected, it is likely that many of the doctrines existing within this society will be at odds with justice as fairness. Some may be entirely illiberal –

[3] A complete explication of the necessary revisions is beyond the scope of this chapter. This is further explored in Chapters 8 and 9.

fascistic or theocratic – while others may be broadly liberal but challenge Rawls' liberalism – libertarianism or utilitarianism, for example. If the principles of justice are to uphold the constitution, the majority of people must be fully autonomous citizens who are motivated to endorse these principles. Respecting people's freedom of conscience, it must be imagined that people are in control of their thoughts and acts and can ensure they maintain a commitment to justice as fairness. This is to make large assumptions about a person's moral psychology.

As stated earlier, this is why self-respect is important for Rawls. It is because people have self-respect that they are motivated to act in ways appropriate to Rawls' well-ordered society. There are two problems here. First, this form of moral psychology, on which self-respect is dependent, will only develop in particular societal contexts. If hard determinism is considered true, we cannot rely on the individual to develop her own moral psychology; the moral psychology she comes to hold will be beyond her control. Second, self-respect requires a world in which we are able to freely choose our actions. In the absence of this world, we would have no justification for thinking that we can freely choose our own ends, and that we are in full control of the processes leading to our choosing these ends. A view of the universe premised on hard determinism would entail the realisation that the physical laws of causation determined our life plans before our births (Van Inwagen and Griffiths 1985), and their virtue and success, or lack thereof, were not due to our freely choosing to be virtuous characters. It is due to events beyond our control that we either become a virtuous person who realises a successful plan of life, or a scoundrel whose way of life undermines society. Neither pride nor shame are justified in response. We could not, therefore, increase our sense of self-respect through imagining that these free choices had enabled us to become such persons. The crucial aspect on which self-respect is founded would thus be lost, and with it, the primary motivation to endorse justice as fairness.

To conclude, whereas full autonomy is not necessarily dependent on the truth of free will, if we are to be considered rationally autonomous, we must be thought of as morally responsibly agents possessing free will. Rawls' conception of moral psychology supports this view of citizens. Without this being the case, we would need to revise how we imagine citizens can become fully autonomous. As I argued in Chapter 3, without fully autonomous citizens, political liberalism would be unrealisable. Rawls places a large demand on the psychological makeup of citizens. Citizens must possess a large degree of control over their thoughts and actions, ensuring their conceptions of the good align with the demands of justice of fairness, and that they are motivated to endorse justice as fairness. If it is not through the motivation supplied by Rawls' theory of moral psychology that people come to endorse justice as fairness, and thus to be fully autonomous, the way in which Rawls imagines we are to realise full autonomy must be reassessed.

5.6 Conclusion

Throughout his work on political liberalism, Rawls tried to avoid metaphysical commitments, attempting to formulate certain aspects of Kantian moral philosophy without their metaphysical components, within a "reasonable empiricist framework" (Rawls 2005, p.285). However, in the original position, though we are supposed to discard all knowledge of ourselves, we retain a particular conception of the self. This conception of the self is one that has been developed in a modern liberal democratic society. It is built upon the acceptance of certain values as being inherently true. We develop our self-respect through the realisation that we are virtuous people who commit noble actions: actions that support the principles of justice and allow us to realise our rational plans of life. Replacing the Kantian noumenal realm with the original position, Rawls still requires us to accept certain values as being true from the perspective of the original position, as Kant imagines universal moral laws can be understood from the perspective of the noumena. One of these values is the truth of free will; without this truth, our virtuous characters and actions are not our own creations for which we can feel responsible.

Without the acceptance of our free will in the original position, we cannot accept responsibility for our conception of the good. In the universe of hard determinism, it would be realised that our conceptions of the good were not a result of our own free choices; as the way in which we conceived the good life was conditioned by factors prior to our births, we could not be viewed as being in full control of the conception of the good we came to hold. Thus, Rawls' conception of rational autonomy is undermined. Furthermore, though full autonomy faces no direct threat from this argument, we become fully autonomous through our capacity for rational autonomy. That is, we become fully autonomous through realising that the principles of justice, and our acting in support of them, enhances our own good. If rational autonomy is undermined, we would need to reconsider how people are expected to become fully autonomous. Thus, if Rawls' formulation of autonomy is to work in the way he intends, there must be an acceptance of the truth of free will. If hard determinism is true, the development of a person's moral psychology is beyond her control. This cannot be the case if we are to realise our rational and full autonomy in the way Rawls envisages.

Chapter 6 – How the Brain Works and what this means for Free Will

6.1 Introduction

Benjamin Libet's work in the 1980s introduced a neural perspective into debate on free will. His work appeared to show that our conscious awareness of our intentions to act followed behind our brain's decision. This suggested that conscious control of our actions is limited. By the time we become aware of our intention to act, we can only decide to veto this action. We can never consciously initiate an intention to act, according to Libet. Though Libet did not entirely discard the notion of free will, as he thought that some kind of conscious control over action was maintained through the veto function, his work nevertheless questioned how much conscious control we have over our actions. Furthermore, Libet precipitated debate within both neuroscience and philosophy on the empirical evidence for free will. The supposed philosophical implications of Libet's work have been questioned by philosophers of the mind, while others working in neuroscience have further explored the way in which our brain's control over our thoughts and actions may bypass our conscious will.

As previously discussed, Rawls gives priority to freedom in *Political Liberalism*, assuming that, with their freedom secured, citizens will be able to formulate a rational plan of life and live accordingly. From the empirical evidence and philosophical implications of neuroscience, I question whether this assumption is plausible. The purpose of this chapter is to assess the empirical data, while the philosophical implications are assessed in Chapter 7. Commenting on the significance of their work, neuroscientists often make philosophical speculations. These will be stated here though more expansive discussion of the philosophical implications of this work is found in the following chapter.

There are two items of fundamental importance to which I wish to draw attention. First, the lack of a centre of conscious control. There is no part of the brain that acts as 'control centre'. Second, neural activity does not occur in linear progressions between the initial conscious cause and the resulting action. Instead, neural activity occurs in deterministic cycles. These two aspects are of philosophical relevance and are further assessed in Chapter 7.

In 6.2, I begin by explaining the basic workings of the brain, and how the various regions of the brain relate to our thoughts and actions. This is followed in 6.3 with an explanation of Libet's work and how his findings challenged conceptions of free will. Contemporary research on the subject in neuroscience will then be explored in 6.4. Finally, I look at findings in social neuroscience in 6.5, and consider how they may affect how we think about free will. Rather than think about free will in relation to the individual agent, social neuroscience looks instead to the ways in which social factors influence neural activity.

6.2 The Anatomy and Workings of the Brain

Beginning with Benjamin Libet's work, research in neuroscience has suggested that our conscious will has little control over our actions. Before explaining how neuroscientists have examined the relationship between the brain and free will, it will be helpful to have a basic understanding of the workings of the brain. The brain and the nervous system are vastly complex; to give a complete understanding of their workings would go beyond the purposes of this chapter. Instead, I aim to provide an explanation of all the main regions of the brain, drawing particular attention to how they relate to the ways in which we think and act. Beginning at the spinal cord, I map the brain between the brain stem and the cerebrum. The function of neurones is then explained. A glossary of key terms is provided at the end of the thesis.

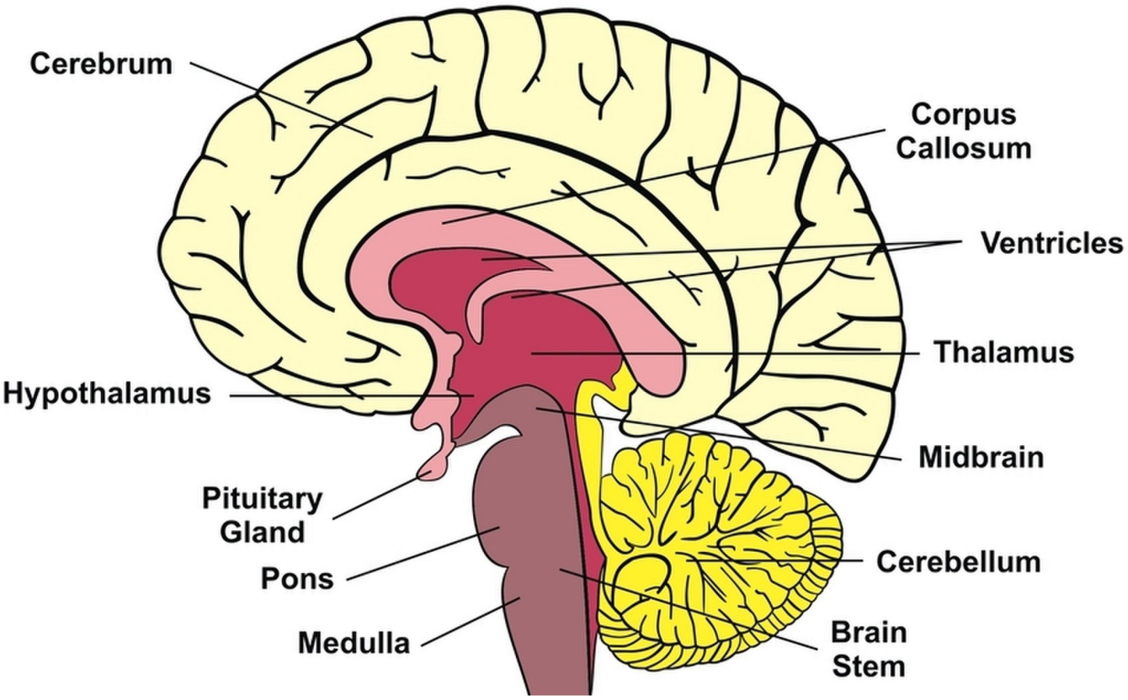

Figure A: Cross section of the human brain (Zimmerman 2017)

The spinal cord and the brain are connected by the *brain stem* (Thompson 2000, pp.14-15), as shown at the base of Figure A. Consisting of the medulla and pons, the brain stem allows for the transfer of nerves between the spine and the brain. The medulla follows from the spinal cord, and connects the brain with the spine, while the pons leads toward the *cerebellum* sitting at the back of the brain (Thompson 2000, pp.15-16), where motor control is regulated. Information is received in the cerebellum from the spinal cord and brain stem, along with other sensory inputs, including the cerebral cortex (Longstaff 2005, pp.254-256). Our conscious awareness of our bodily movements derives from this information conveyed to the cerebellum. From the cerebellum, information is then relayed to other parts of the brain in order to guide movement.

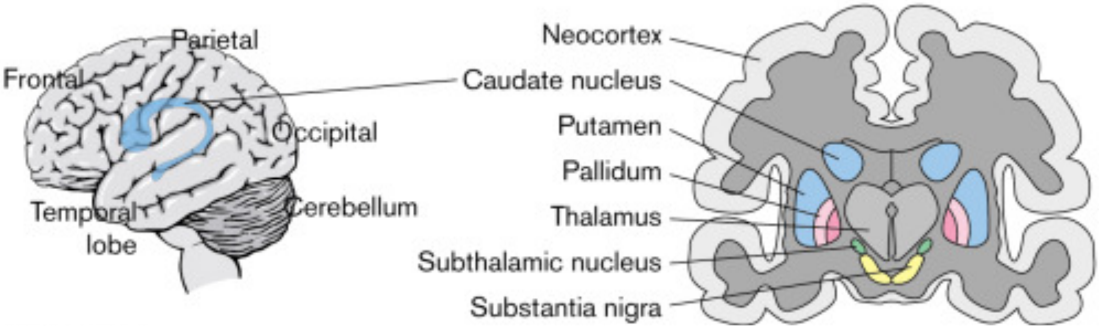

Figure B: Location of the basal ganglia (Graybiel 2000)

Above the cerebellum is the *basal ganglia*, the location of which is shown in Figure B, further parts of the brain concerned with bodily movement (Thompson 2000, pp.18-19). The basal ganglia consist of a large group of nuclei in the centre of the cerebral hemispheres. There are two structures constituting the basal ganglia: the dorsal pallidum (globus pallidus), positioned below, and the caudate nucleus and putamen, located above. Together with the nucleus acumbens, the caudate nucleus and putamen combine to form the *striatum*, important for transmitting chemicals such as dopamine (Kingsley 2000, pp.285-289), along with allowing for communication between neurones (to be discussed shortly). Dopamine's release also involves the basal ganglia, via the *substantia nigra* that produces it (Kingsley 2000, pp.131-133). The basal ganglia were also thought to contribute to the regulation of voluntary movements, and the inhibition of involuntary actions (Mink 2003). However, the division of the brain into systems dealing with voluntary movements and others dealing with those that are involuntary had largely been rejected by the early twenty-first century (Roth 2003a). As will be discussed, those such as

Gerhard Roth argue that there is interplay between the different systems, and that the basal ganglia are involved in more tasks than originally thought, both voluntary and involuntary.

The *thalamus* is next to the basal ganglia. It is another large group of nuclei, consisting of two ovoids, one within each hemisphere (Thompson 2000, p.16). It is involved in the relaying of sensory information. Connecting the thalamus to the brain stem and spinal cord is the midbrain. At this junction between the thalamus and the midbrain sits the *hypothalamus* (Thompson 2000, pp.15-16). The hypothalamus is a group of small nuclei. According to Richard F. Thompson (2000, pp.15-16.), it has a large amount of control over the body, due to the power it has over the pituitary gland. The hypothalamus is the control centre for the endocrine system, the system concerned with the release of hormones throughout the body, which also encompasses the pituitary gland through a feedback mechanism (Musumeci et al 2015, pp.357-358). This system controls human physiological processes such as growth, metabolism, and fertility.

Another part of the brain is known as the *limbic system*, though this consists of various sections located across the brain. The *amygdala, hippocampus, limbic cortex*, and *septal area* form the main parts of the limbic system (Thompson 2000, pp.17-18). The limbic system performs various functions. It is one of the earliest parts of the brain to have evolved, and its original role was to formulate an organism's response to sensory stimuli, particularly in relation to smell, though as it has evolved, it has come to perform various other functions. Recent work has uncovered the possibility of there being several limbic systems, with the main parts of the central limbic system forming different networks. One such network involves the amygdala and is concerned with emotion, while the hippocampus is part of a network that relates to memory and learning (Rolls 2015). Also involved with the limbic system is the *cingulate cortex*, which, along with the amygdala, regulates emotional responses (Bush et al 2000).

All of the aforementioned regions are covered by the *cerebrum*, the largest part of the brain, located at the top and reaching from the front to the back. As seen in Figure B, the cerebrum is divided into four lobes: frontal, parietal, occipital, and temporal (Graybiel 2000; Salat 2004). The grooves and folds on the surface of the lobes are named gyri (gyrus) and sulci (sulcus) (Kingsley 2000, p.7). Contained in the cerebrum is the *cerebral cortex*. Thompson writes that the cerebral cortex is "what makes human beings what they are" (Thompson 2000, p.19). Our consciousness, ability to reason and imagine, capacity for language, along with our senses and motor skills are thought to belong to the cerebral cortex (although, as will be discussed later, the importance with which the cerebral cortex is traditionally considered is now questioned). An important part of the cerebral cortex is the *prefrontal cortex* positioned within the frontal lobe that is concerned with memory and learning. Another part is Wernicke's area, lying at the end of the Sylvian fissure that separates the frontal and temporal lobes (Thompson 2000, p.441). Within Wernicke's area are the *posterior superior temporal gyrus* and *sulcus*, which are important in regard to speech (Friederici et al 2009). The cerebral cortex is also important in action and movement. When we decide to act, it was traditionally thought that the initiation of the act is in the cerebral cortex (Eccles 1972, pp.108-109). Cells are then fired into the *motor cortex*, found within the cerebral cortex (Eccles 1972, p.105), which prepares the body to perform the act. However, as is discussed later in this chapter, this understanding of the brain, on which there are neural regions where events are initiated, is challenged by contemporary neuroscience.

The motor cortex is divided into three further subsections: the primary motor cortex, premotor cortex, and the supplementary motor area (both the pre-supplementary motor area and supplementary area proper) (Roth 2003b, p.116; Ward 2015, pp.168-172). Voluntary movements are often understood to involve the primary motor cortex, which sits in the middle of the cerebral cortex, across both hemispheres. The right hemisphere controls the left-hand side of the body, and the left hemisphere the right. Modulation of actions is performed by the premotor cortex, while the supplementary motor area regulates well-learned actions that do not require monitoring of the environment, such as the playing of a musical instrument (Ward 2015, pp.168-172). Sometimes considered as part of the motor cortex are the posterior parietal cortex (Roth 2003a, p.111) and the primary somatosensory cortex (Kingsley 2000, p.84), which relays sensory information to the thalamus.

Within all of these regions and sections of the brain are cells called *neurones*. Neurones are contained within the brain and nervous system, and there are estimated to be between 300 and 500 billion neurones within the human body (Longstaff 2005, p.5). From the neurone cell, nerve fibres called *dendrites* branch upwards and sideways, while *axons* grow down (Eccles 1972, pp.4-6). It is through these nerve fibres that information is transferred between neurones. The transferring of this information enables the activity in our brains that allows for our thoughts and actions (Eccles 1972, pp.9-10).

Neurones communicate with each other via synapses. When a neurone communicates with another neurone, it initiates what is called an *action potential* (Postle 2015, p.34). A neurone is said to have an action potential when it is electrically stimulated; neurones not undergoing stimulation are said to have a resting potential (Longstaff 2005, p.33). The action potential occurs when a neurone has been depolarised. Depolarisation is the process whereby the voltage of a neurone has been sufficiently increased for the opening of the neurone's channels, enabling an influx of sodium ions. For the action potential to occur, the part of the neurone known as the axon hillock, the branch leading down to the axon, must be depolarised. Further down the axon are more clusters of sodium ions. Experiencing the depolarisation that has occurred in the axon hillock, these clusters open the part of the channel they inhabit. This process repeats itself continuing down to the dendrite of the next neurone, and the action potential is then transmitted between neurones. In the postsynaptic neurone, the neurone receiving the information encoded in the action potential, there can be many possible effects (Postle 2015, p.35). There will be at least a little amount of depolarisation in this neurone; with a sufficient amount of depolarisation, a further action potential may be enabled in the postsynaptic neurone.

Three types of motor acts are identified: autonomous motor functions, reflexes, and voluntary motor acts (Kingsley 2000, p.209). Autonomous motor functions are acts such as the beating of the heart, which continue without conscious awareness or intervention, and are independent of external stimuli. Reflexes happen automatically in response to stimuli, such as being startled by an unexpected loud noise. Finally, voluntary motor acts are those in which we are understood to be consciously involved. There are two motor systems within the brain concerned with regulating movement, the pyramidal system and the extrapyramidal motor system (Thompson 2000, p.19), though it should be noted that those such as Gerhard Roth no longer view there being a sharp distinction between the two (Roth 2003a, pp.111-112). The pyramidal system descends down to the spine from the motor cortex and includes sections relating to various parts of the body (Eccles 1972, pp.105-107). Traditionally, it was thought that the pyramidal system was concerned with voluntary actions. The pyramidal system includes the whole of the aforementioned motor cortex along with the pre-frontal cortex (Roth 2003a, p.111). Pyramidal cells are fired from the motor cortex down this chain towards the muscle attached to the body part we wish to move (Eccles 1972, pp.105-107). The extrapyramidal motor system, though similarly originating in the motor cortex, bypasses other parts of the pyramidal system and was thought to regulate involuntary actions (Thompson 2000, p.308; Whitty et al 2008, p.416). It is largely comprised of the basal ganglia, brainstem, and motor centres in the spine (Roth 2003a, pp.111-112). Despite Roth's argument that there is not such a strong distinction between systems in the brain dealing with voluntary movements as opposed to those relating to involuntary movements, this terminology is still often used in the literature, and a familiarity with the distinction is useful.

The above provides a basic illustration of the workings of the human brain. From the spine up to the cerebrum, each part of the brain is involved in different functions in relation to our thoughts and actions. The medulla and pons allow for the transmission of information between the brain and the spine. Motor control is regulated by the cerebellum. The basal ganglia allow for our body movements and are involved in the regulation of involuntary movements. Sensory information is conveyed by the thalamus. Though Roth contests its significance as a command centre (Roth 2003b, p.115), Thompson argues that the cerebral cortex allows for us to think as humans do: to reason and imagine, to learn languages and other forms of communication, and to interpret our environment (Thompson 2000, p.19). Within all these regions of the brain are neurones, working to transmit information across the brain and body and enabling us to function as we do.

6.3 The Readiness Potential and Conscious Awareness

In this section, I explain the *readiness potential* before moving on to Libet's experiment, which examined the relationship between the readiness potential and our conscious awareness. I then explore the ramifications of Libet's experiments for the subject of free will. The readiness potential was discovered in an experiment by Hans Helmut Kornhuber and Lüder Deecke in 1965 (Kornhuber & Deecke 1965; Libet 1999, p.49). Before the performance of any action, electrical activity in the brain – the voltage increases that allow for depolarisation, which further leads to the communication of information between neurones – occurs in relation to the intention to act. This electrical activity was named the readiness potential by Kornhuber and Deecke. In his own experiments, Benjamin Libet further explored the implications the readiness potential held for how we commonly think about free will.

Sixteen subjects were asked to lie in an electrically shielded room and to sometimes flex their right index fingers, and at other times to flex their right arms (Deecke et al 1969, p.159). They were asked to fix their gaze and refrain from making other movements. Electrodes were then placed on the subject's skin across his or her brain. By looking at electroencephalographic data obtained from the experiments through reverse computation, Kornhuber and Deecke discovered that the readiness potential almost always occurred prior to hand and foot movements (Kornhuber & Deecke 1965; Gomes 1999, pp.62-63). Though on average it begins at 850 milliseconds prior to the act (Deecke et al 1969, p.163), it can happen up to 1.5 seconds beforehand (Gomes 1999, p.62).

The readiness potential is not to be confused with the action potential described above. While the action potential occurs within each neurone as it prepares to communicate with other neurones, the readiness potential refers to a larger group of activity involving many neurones. Nevertheless, within the readiness potential, the transfer of information up to the point of the muscle moving the body occurs through action potentials (Keller & Heckhausen 1990, p.359). There has been debate within neuroscience on what is happening during the readiness potential and where (Böcker et al 1994, pp.275-276). From experiments on macaque monkeys, it was thought that activity in the pre-motor cortex, primary motor cortex, and the somatosensory cortex occurred in succession (Sasaki & Gemba 1991). Others argued that the readiness potential activity can be solely attributed to the primary motor cortex (Neshige et al 1988). Praamstra et al conclude that it is likely that the supplementary motor area plays a small role in the readiness potential (Praamstra et al 1996, p.476). There is agreement, however, that the readiness potential is initiated in the motor cortex. It is also understood that there are higher levels of the readiness potential in the performance of voluntary actions. When actions are initiated involuntarily or through external stimuli, there are fewer signs of the readiness potential (Praamstra et al 1996, p.468).

Libet conducted his own experiments to determine the relationship between the readiness potential and our conscious will. He had his subjects look at a clock face that had been altered so that the second hand moved 25 times faster than normal (Libet 1999, pp.49-51), in order to account for milliseconds. Subjects were asked to flick their wrists at a time of their choosing. When the subject first felt the urge to act, he or she was asked to note the position of the hand on the clock face. The experiment revealed that while the readiness potential registered at 500 milliseconds prior to the act (Libet's finding here contrasts with the 1.5 seconds to 850 milliseconds observed by Deecke et al), the conscious awareness of the urge to act was apparent at 200 milliseconds before the act. This was evident, with slight variations, across all of Libet's results. Thus, the readiness potential always precedes our conscious awareness of our desire to act.

By the time we become aware of our intention to act, Libet thought that we could still veto this intention (Libet 1999, pp.51-52). At times, Libet's subjects noted an urge to act that they suppressed. This was shown in the activity of the brain too. Between 100 to 200 milliseconds before an act, by which time the individual is aware of the urge to act, the act can be cancelled. However, we have only a short time to exercise this veto. The spinal nerve cells are activated by the motor cortex in the final 50 milliseconds before the act, preparing the body for action, at which point there is no longer an opportunity *not* to act. It is quite possible, however, that the veto itself has unconscious origins (Velmans 1991; Libet 1999, pp.52-53). While Max Velmans considered even an unconsciously originated veto to be a sufficient condition for free will, Libet thought that in order for actions to be considered the result of a person's free will, the veto function had to restore some kind of role for conscious control. Without this, even the acts of someone experiencing an epileptic seizure would have to be considered acts of free will (Libet 1999, p.52). This was unacceptable for Libet, who thought that we must be able to distinguish voluntary movements from the type of involuntary motor actions that characterise seizures.

For Libet, then, the veto had to salvage a role for the conscious will in the performance of our actions for acts of free will. What was important for Libet was the idea of *awareness* (Libet 1999, p.53). While we may not consciously initiate the process that occurs between the brain's intention to act and the performance of the act itself, Libet thought it possible that we are *aware* of this whole process. Libet argued that there was no logical imperative in any mind-brain theory that required our conscious control to be preceded by neural activity. Hence the veto function, as Libet posited it, could exist independently of prior neural states. Once we become aware of an intention to act, this awareness, in conjunction with the veto function, allows for the possibility of our free will. Essentially, our conscious awareness is watching over the entirety of the process, and before the execution of an act, can intervene to halt the process.

This was enough, for Libet, to argue that we do have free will, and a sufficient level of free will to uphold many ethical systems in religious and philosophical thought. As Libet writes, "most of the Ten Commandments are 'do not' orders" (Libet 1999, p.54). We have enough control over our actions to refrain from an action we know to be considered ethically or morally wrong. What we do not have, however, is control over our thoughts. Thus, while it would be acceptable to hold someone responsible for immoral actions, we could not hold people responsible for their thoughts.

Though some have used Libet's experiments to support a hard-deterministic stance on free will (Harris 2012), or to argue against any conscious control of action (Wegner 2002, pp.52-55), Libet himself argued that his findings did not rule out free will's truth (Libet 1999, pp.55-57). Libet also thought that the truth of free will was essential for life to be worthwhile living. While he thought determinism was useful when examining the natural world, Libet saw nothing to definitively conclude that the mind worked entirely deterministically. Therefore, he argued against assuming a deterministic stance in relation to free will. Instead, Libet thought that at the conscious level, we are continually aware of our actions, and through the veto function, we can consciously select the actions we decide upon to realise our ends. On this view, conscious awareness is not a phenomenon that follows behind our actions, nor are the justificatory reasons we give for our actions decided on after the act itself.

6.4 Understanding of Free Will in Contemporary Neuroscience

Since Libet's work, some philosophers and neuroscientists have disputed Libet's interpretation of the data, while others looked to advance on his findings. The philosophical responses to findings in neuroscience will be considered in the following chapter. I now turn to the response from neuroscience, and the ways in which free will is considered within contemporary neuroscience. Beginning with Patrick Haggard's work, I then turn to that of Gerhard Roth, before finally examining alternative perspectives in neuroscience.

One of the primary neuroscientists to have advanced the relationship between neuroscience and free will is Patrick Haggard. Haggard has conducted further research into the neural activity behind actions that actors perceive as being voluntary. Modern neuroscience, according to Haggard, rejects a dualistic conception in which the conscious mind, soul, or will dictates to the brain and body what it must do (Haggard 2008, p.944). Rather, there are specific parts of the brain that relate to the performance of voluntary actions. Following Libet, Haggard does not suggest that our awareness merely follows neural and bodily activity, and our explanations for our behaviour are not constructed after we have already acted (Haggard 2008, p.942). This can be demonstrated through experiments in which electrodes artificially stimulate motor areas of the brain. When the pre-supplementary motor area is stimulated, subjects report feeling an urge to move, though as the body has not actually moved, there is no need justify our action. Awareness was a key part of the whole process between intention and action for Libet; Haggard, similarly, finds that our awareness runs in tandem with the process.

Thus, Haggard posits that specific areas of the brain are concerned with voluntary actions, and the feelings we experience in relation to voluntary actions are not only a construct to justify our behaviour. Haggard has further identified the regions of the brain involved in the performance of actions. Though the primary motor cortex is generally considered to be the area from which commands originate, there are several inputs feeding into the primary motor cortex (Haggard 2008, p.936). What Haggard describes as a key input derives from both the basal ganglia and the prefrontal cortex, from where it travels to the pre-supplementary motor area, before arriving at the primary motor cortex. As previously mentioned, the pre-supplementary motor area has been identified as a possible location from which the readiness potential originates (Yazawa et al 2000; Shibasaki & Hallett 2006; Haggard 2008, p.936). Haggard finds this problematic for two reasons. Firstly, the readiness potential is generally considered as being the electrical charge that occurs in the milliseconds prior to action, whereas Haggard states that there is research suggesting it begins at a much earlier stage (Soon et al 2008; Haggard 2008, p.936). Secondly, activity in the pre-supplementary motor area must itself have a cause. According to Haggard, neural activity works in loops rather than in a linearity of causes extending back to an "uncaused cause" such as the conscious will (Haggard 2008, p.936).

Within this key input, Haggard views the basal ganglia as having an important role. People with Parkinson's Disease perform uncontrollable, involuntary actions. During these actions, activity in the loop between the basal ganglia and pre-supplementary motor area is reduced (Haggard 2008, p.936). Furthermore, as the basal ganglia enable the release of dopamine, basal ganglia activity leads us to modify our behaviour in order to experience reward. Haggard suggests that this is how organisms learn to interact with both their historical and current environment. Voluntary action is a result of this process of interaction, not an isolated event localised within the conscious will of an agent.

Gerhard Roth has also conducted further research into the nature of voluntary actions. Central to Roth's work is a rejection of the cerebral cortex as a command centre from which decisions flow (Roth 2003b, p.115). Roth also views the basal ganglia as being important within the execution of voluntary movements. There are two basic systems involved in these movements, according to Roth. One includes the pre-motor cortex, the primary motor cortex, and the cerebellum (Roth 2003b, pp.120-121). This system is behind actions that are learned and well-practised, such as playing a musical instrument, and the cerebellum ensures the sequencing and smooth running of these actions. Little thought is required for their execution. Actions that necessitate further deliberation are regulated by another system. This system includes the basal ganglia (Roth 2003b, pp.121-123). When we have to make a decision, no matter how trivial, activity in the basal ganglia enables the decision to be made, whether we make that decision consciously or unconsciously.

The basal ganglia are connected to the cerebral cortex via three loops (Roth 2003b, pp.123-126). The first loop regulates planning and preparation, the second loop relates to the execution of actions, and the third loop relays cognitive, emotional, and motivational information. Within these loops are further pathways, some of which are excitatory and others inhibitory. The striatum has an inhibitory function over the substantia nigra, which then inhibit the thalamus, which performs an excitatory function. Due to the dopaminergic inputs from the basal ganglia, via the striatum and substantia nigra, the loops between the basal ganglia and the cerebral cortex can regulate behaviour in accordance with what is known to obtain reward. When we make what we perceive to be a voluntary decision, the work of the basal ganglia has allowed for this decision to be made (Roth 2003b, pp.123-126.).

Roth also argues that the basal ganglia have an important role in the readiness potential. He writes that the readiness potential exists through two components (Roth 2003b, pp.126-127). One is the *symmetrical* component, involving both parts of the supplementary motor area, and beginning 1-2 seconds before the act. The other is the *lateralised* component, which can be found in the cerebral hemisphere opposite to the body part to be moved (due to the right hemisphere controlling the left side of the body and vice versa). This begins 700-500 milliseconds before an action. Roth argues that the symmetrical readiness potential – though involving the activity of neurones in the supplementary motor area, before moving to the pre-motor cortex and primary motor cortex – is directed by the basal ganglia via the thalamic relay. However, Roth is aware that the origination of the dopaminergic processes that allow for these functions does not lie with the basal ganglia itself. Instead, the limbic system, with its storing of memories and emotions, is what triggers the release of dopamine, though it does so at a largely unconscious level (Roth 2003b, pp.127-129).

The basal ganglia, then, allow for the selection of the actions we make. Roth states that when we decide what action to make, the basal ganglia – through their access to memory – reflect on whether it is an appropriate action, and whether, based on past experience, it is more suitable than other actions (Roth 2003b, pp.129-130). All of this occurs, however, at a level that is, in the main, unconscious. Roth's work further diminishes the claim that we may, in some way, still have conscious control over our actions, a position that Libet maintained. Rather, our feeling of freedom – in which we feel as though we are consciously and voluntarily choosing our own thoughts and actions – arises from our conscious intentions cohering with our unconscious plans, according to Roth (Roth 2003b, pp.129-130.).

Neither Haggard nor Roth explicitly endorse hard determinism, though in both of their work, the activity of the brain is seen to occur deterministically, and without conscious intervention. Though Haggard is concerned with what constitutes voluntary action from a neural perspective, what is *voluntary* in the neural sense is not necessarily *free* in the metaphysical sense. Activity in certain regions of the brain may indicate that an action is perceived as being voluntary, yet this does not mean that it was freely and consciously chosen by the individual. As Schopenhauer wrote, "you can do what you will, but in any given moment of your life you can *will* only one definite thing and absolutely nothing other than that one thing" (Schopenhauer 1839, p.24). Our will may be able to guide our action, but the will cannot exert control over itself through freely choosing its own ends. Hence there is a difference between something perceived as being voluntary and something being free.

While there is a range of opinion in neuroscience on the nature of free will, and not all neuroscientists find the inner workings of the brain to be deterministic (Brembs 2010),[1]

[1] It should be noted that this is not because some neuroscientists maintain that there are necessarily uncaused causes within the brain, and that the conscious will can trigger a chain of activity between such a cause and an action. Instead, neuroscientists do not rule out indeterminism because of findings within quantum physics.

following Libet's work, most neuroscientists commenting on the subject have looked at the deterministic ways in which the brain works, and how this occurs at an unconscious level. Though Haggard and Roth have not outright denied that was have free will, their work has further revealed how little influence the conscious will has over human action. While Libet thought that the veto function could be consciously controlled, Haggard and Roth's work suggests that much of the neural activity occurring temporally prior to an act is unconsciously performed. Whether we decide to follow through with the initiation of an act or veto it, our conscious will has little involvement according to Haggard and Roth.

To conclude, as stated in the Introduction, there are two important aspects found within the work of Haggard and Roth. First, neural activity occurs circularly rather than linearly, and does not extend back to an "uncaused cause". Second, there is no centre from where all of this activity is controlled. Different neural regions perform certain activities, without the need to be linked back to some kind of control centre. Both of these findings have important philosophical implications which are explored in the following chapter.

6.5 Free Will in Social Neuroscience

Research in social neuroscience has further contributed to our understanding of the role of consciousness in decision-making. Rather than examine the individual brain, the social neuroscientist thinks of brains as existing in networks. As we are thought to be social animals, it is the social network in which we exist that exerts the most powerful influence over the brain's development. It is therefore inadequate to consider human brains in isolation, according to social neuroscience. In this section, I examine how work in social neuroscience could affect the way in which we consider free will.

The social neuroscientists John Cacioppo and Gary Berntson posit that consciousness would be epiphenomenal *if* it existed in isolation. Their argument for this comes in two parts (Cacioppo & Berntson 2012, pp.41-42). Firstly, consciousness is the result of temporally prior brain states. A state of consciousness, according to Cacioppo and Berntson, is entirely predicted by previous activity in the brain. Secondly, consciousness has no control over subsequent brain states. If one brain state at a particular time causes another brain state, then consciousness cannot also be the cause of the second brain state; if there is sufficient causation at the physical level, there cannot also be causation elsewhere. Consciousness is thus posited as being epiphenomenal; it is a by-product of neural activity.

Within the singular brain, consciousness would be functionless if it were epiphenomenal. However, according to Cacioppo and Berntson, this is not the case. We are social animals, and our brains are thus designed to work in conjunction with the brains of others (Cacioppo & Berntson 2012, pp.44-47). Our very survival – as both individuals and as a species – depends on us being able to cooperate with others, and we therefore learn how to communicate with others from a young age. This prepares us to react to the behaviour of others. Brain states are, then, not only the result of prior internal brain states, they are also the product of our interactions with others. Interactions change our own brain states. In this context, Cacioppo and Berntson posit that consciousness – rather than being epiphenomenal and functionless – has a social function. Our conscious states, formed by our beliefs and intentions, influence the brain states of other people (Cacioppo & Berntson 2012, pp.47-48).

Cacioppo and Berntson recognise that this could be characterised as the brain states of some influencing the brain states of others, with no role for consciousness (Cacioppo & Berntson 2012, pp.47-48). However, they counter this by comparing our conscious states to the output display on a computer screen. Without a user, the display will remain static and will not affect the operations that first produced the output. When someone uses the display to operate the computer, the display itself influences the actions of the user, in turn influencing the future states of the computer. Our conscious states similarly influence the behaviour of others, and in turn, affect our own future brain states. Because we cannot accurately predict the behaviour of others, the mechanistic operations of our own brain states are interrupted by the unexpected behaviours of those around us. Consciousness, thus, through its social function, serves to disrupt the internal determinism of the brain. Social processes, therefore, play a large role in determining human behaviour.

Our capacities for decision-making, then, depend on our socialisation. When analysing human behaviour, rather than look to the individual agent, the social neuroscientist observes the social environment. Thus, Yoder and Decety argue that morality, as a guide for human behaviour, is itself a product of the influence of social relations on the brain (Yoder & Decety 2018). Two arguments support this theory. Firstly, morality has social utility; when humans behave according to rules that lead to mutually beneficial outcomes, social stability can be better ensured (Yoder & Decety 2018, p.283). Secondly, as Hume argued, our moral reasons follow our moral emotions (Hume 1738; Yoder & Decety, 2018, p.284). We construct arguments to justify our initial emotional reactions to events, a claim also supported by Haidt, as noted in Chapter 3 (Haidt 2007). These emotions also serve a social function. Empathy and shame, along with other emotions, encourage us to obey the rules of our groups, and to care for others, ensuring the group's stability and survival. These emotions have a neural basis in the amygdala of the limbic system. When we view actions which we perceive as having harmful consequences, the amygdala, in connection with other regions of the brain, modulates our response (Yoder & Decety 2018, pp.285-286). The posterior superior temporal sulcus, posterior cingulate cortex, and the medial prefrontal cortex are regions of the brain concerned with the interpretation of the beliefs and intentions of others (Yoder & Decety 2018). If a harmful act is thought to be accidental, these regions regulate the response of the amygdala to the dorsolateral prefrontal cortex. We are less likely to wish to punish someone guilty of accidental wrongdoing because of this communication between the amygdala and the dorsolateral prefrontal cortex.

The neural bases that support these social and moral functions are thought by some in social psychology and neuroscience to be inherent aspects of the human brain. While it was traditionally thought that empathy was a learned trait, psychologists now argue that the capacity for empathy is found in very young children (McDonald & Messinger 2011). Children show signs of distress when witnessing stress in others, and a desire to comfort those who are suffering. One hypothesis for this is the existence of mirror neurones (McDonald & Messinger 2011.). Mirror neurones have been observed within macaque monkeys, though their existence within humans is disputed. The mirror neurone prepares the organism to relate to the experiences of other organisms and act in accordance. Despite their existence within humans being questionable, some neuroscientists posit that it is the mirror neurone that prepares us to react to the behaviour of others (Ferrari & Coudé 2018, pp.68-70). The activity of mirror neurones occurs during everyday actions and reflexes, such as yawning in response to seeing others yawning, a behaviour that has been observed in monkeys as well as humans. Ferrari and Coudé argue that humans emulate the emotions of others due to the presence of mirror neurones. Thus, when we see others smile, we smile in return, a phenomenon termed "emotional contagion" (Ferrari & Coudé 2018, p.73). Whether or not it is the presence of mirror neurones that prepare humans for empathic behaviour, it is clear we are inherently predisposed to respond to the emotional states and behaviour of others.

Though there is a neural basis that prepares us for empathic behaviour, its development is dependent on how we are raised. Those whose parents are more attentive are likely to have a stronger conscience later in life (McDonald & Messinger 2011). Our brains may hold an intrinsic capacity for empathy, but this capacity will not develop uniformly from person to person. Instead, development is conditional on environment. The processes influencing this development occur outside of the individual agent's conscious will, both at the neural level and at the familial and societal levels. This is not only true of the development of empathy. Our feelings of shame and guilt also develop in response to parenting styles (Parisette-Sparks et al 2017; Ruckstaetter et al 2017). Emotions that motivate us to follow particular moral laws only develop under particular circumstances. Understandings of morality that hold in the individual's family and community will, then, influence the brain from a very early age. The brain has thus been formed and moulded by a myriad of external forces prior to adulthood. While we may be inherently social animals, the precise quality of our sociability is dependent on our environment.

While Libet retained some space for the conscious will to have control over action via the veto function, Cacioppo and Berntson also find consciousness to serve a function, though they offer a different formulation. Instead of our awareness of the process between intention and action affording us the ability to decide on appropriate actions, under Cacioppo and Berntson's formulation, we instead come to make decisions through our sociability. There is, however, little sense in which this can be considered free. Our decision-making processes are instead largely guided by our reactions to the behaviour of others, both in the immediate sense, and based on past experience. In the work of Roth, we come to something of a black hole when we seek the origins of intentions. Neural systems mainly work in loops, with one region of the brain influencing another, and with no root to the intention guiding activity. Cacioppo and Berntson offer the possibility that the roots of intentions grow out of our social relations. Furthermore, our senses of morality and emotions, such as shame and guilt, are also dependent on our social environment. Specific parts of the brain are inherently equipped to deal with our emotions and social relations, yet the precise development of the brain will depend on the experiences to which it is exposed. The decisions we make as adults are influenced by both our current social environments, and the social environments in which we were raised.

6.6 Conclusion

Examining the implications of the discovery of the readiness potential, Libet constructed experiments which appeared to show that the readiness potential preceded conscious awareness. From this, Libet concluded that our ability to consciously control our actions is limited, though he did not entirely eschew the idea of free will. Instead, Libet thought we retained the capacity for free will through the ability to veto an impulse to act. We hold an awareness of the entire process between unconscious intention and action, and through this awareness, we are able to consciously veto our thoughts and impulses, maintaining a level of conscious control over actions.

Following on from Libet, some neuroscientists have postulated an even smaller, perhaps non-existent, role for the conscious will. Haggard argues that we do distinguish between voluntary and involuntary actions, with specific parts of the brain being involved in the performance of voluntary actions. However, voluntary acts are not caused by the conscious will, according to Haggard. Neural activity occurs within cycles, and there is no endpoint that can be assigned the 'cause'. Roth has explored the way in which our brains make decisions and found that much of this activity occurs at an unconscious level, with little involvement of the conscious will.

Human behaviour does not spring from neural activity occurring entirely in internal cycles, however. External information must be processed for us to understand our environment and act and respond appropriately. Within social neuroscience, it is our interactions with others that provide the predominant influence on our behaviour. From an early age, we learn to communicate with others, and this interaction shapes the development of our brains. Our decision-making processes, then, rely on our interactions with others. Rather than atomised, individual agents consciously initiating their own actions according to their free will, social neuroscientists find the process from intention to action as being influenced by sociability.

Though there is disagreement within neuroscience itself, few neuroscientists find the libertarian perspective on free will plausible. We could not hold "Ultimate Responsibility" for our ends and actions, as Kane desires us to (Kane 1996), as there is no neural 'end' from which our intentions are triggered. Cacioppo and Berntson, in suggesting that if any 'end' exists, it lies in our social environment, have uncovered a new perspective in our understanding of how humans come to think and act. Our thoughts and actions do not originate in a vacuum, they instead form through processes of social interaction. Instead of looking for the roots of our intentions in the brain itself, as Libet did, social neuroscience suggests we should look to the way in which external events influence the brain. Both at the neural and societal levels, however, there is little freedom for the conscious will of the individual to control events.

Chapter 7 – Philosophical Responses to Neuroscientific Data

7.1 Introduction

The readiness potential, electrical activity in the brain indicating an intention to act, occurs temporally prior to conscious awareness of the urge to act, according to Benjamin Libet. For Libet, this meant that we should reconsider conceptions of free will. This claim precipitated further research in neuroscience examining how the workings of the brain may undermine free will. Neuroscientists such as Patrick Haggard and Gerhard Roth conclude that many functions of the brain that prepare us for action – along with our processes of deliberation –occur deterministically and beyond the limits of conscious thought.

There have been a number of philosophical responses to Libet's work. Many of these responses question precisely how we formulate free will. If we demand – as libertarians do – that if we have free will, determinism must be false, our conception of free will is challenged by the findings of Libet's experiment, along with findings made in neuroscience since Libet's work. These findings show that the brain works largely deterministically; there is little reason to suppose that there is some kind of break within the determinism of neural activity that supports the libertarian conception of free will. It is not immediately clear why compatibilist conceptions are similarly challenged. In this chapter, I assess the different ways in which free will has been formulated in response to work in neuroscience, and the extent to which these formulations resolve the problems neuroscience identifies. Through conducting this assessment, the aim is to establish claims that can be drawn from the empirical data and assessed against Rawls' conception of autonomy.

The fundamental claim I defend in this chapter is that we have little conscious control over our thoughts. This is problematic regardless of our stance on free will. While the libertarian position on free will is difficult to reconcile with the empirical data, compatibilism does not face the same problem. If we imagine that, providing there is a clear lineage between a person's intentions and her acts, then the acts are acts of free will, there is no reason to presume the empirical data challenges our conception of free will. However, this does not resolve all of the concerns raised within neuroscience. If human beings have little conscious control over their thought processes, then the thoughts a person comes to accept or value is largely beyond her control too. While the compatibilist position resolves some of the issues raised by neuroscience, this remains problematic for the stability of the liberal order Rawls' theorises.

I begin in 7.2 by examining the way in which Libet himself formulated free will, assessing the veto function. In 7.3, I look at alternative ways of interpreting the data. Motivations and the reasons that move us to act are examined in 7.4. The idea of the Cartesian theatre – or its non-existence – is examined in 7.5. Following this, in 7.6, I look at the role of social influences and environmental factors in determining the will. The argument across sections 7.2 to 7.5 is largely aimed at undermining the libertarian response to Libet's work (including Libet's own argument for free will). However, in 7.6 and 7.7, I examine possible compatibilist responses to the data, and argue that compatibilism does not resolve some of the concerns raised by the neuroscientific data.

7.2 Libet and the Veto Function

While some such as Daniel Wegner have used Libet's work to support the claim that the conscious will is an illusion (Wegner 2012), it was never Libet's intention to argue against the truth of the conscious will or free will. Instead, Libet, thought that the veto function allowed for our free will. Here, I argue that Libet actually formulated a libertarian conception of free will; the veto function remains undetermined and enables conscious agential control. Subsequent work in neuroscience, however, has shown that the neural processes involved in vetoing are themselves determined, ruling out the existence of an independent cause of action within the mind. For this reason, I argue that Libet's conception of free will is mistaken, providing reasons for endorsing the claim that the veto function is deterministic. The main problem identified by Libet, that the origins of our intentions are beyond our conscious awareness, remains intact; this gives us reason to accept that the origins of our thoughts are prior to conscious awareness.

Libet claims that conscious control over the veto function enables the conscious will to break free of the deterministic laws surrounding it. Libet's claim is, then, supportive of libertarianism. It lends support to the claim that free will is true as determinism is false. We have conscious control over this break with deterministic laws, and this control is what constitutes free will. Libet claimed that the veto function was subject to conscious control. Through this function, we can consciously control the selection of the actions we make, though the intention to act itself is formed unconsciously. The determinism of physical laws would produce the same result as theological determinism, in which our fates were decided by a God, according to Libet (Libet 1999, p.47). If either were true, human beings would be nothing more than automata, and our consciousness only epiphenomena. No sort of determinism is compatible with free will for Libet. Thus, it was essential for Libet that the veto function allowed for conscious control of our actions through offering a break with other deterministic processes.

It is important to note, however, that Libet's findings, and those made since Libet, are not only posing the same problems found within the free will-determinism debate. This is a problem of ordering rather than causation. Whether we live in a deterministic or an indeterministic universe, those who argue for free will may want the conscious will – or our conscious awareness of our will – to precede or be simultaneous with brain activity, rather than follow on from it. Rather than simply repeating the same arguments surrounding the free will and determinism debate, Libet introduces a new idea: the conscious will follows behind neural activity. One could be a compatibilist and still see a problem with this. Libet did not argue for compatibilism, however. Instead, Libet argued as a libertarian; for the truth of free will, we must have independence from deterministic laws in the choices we make.

This is why it was vital for Libet that the veto function allowed for conscious control over our actions. Libet held that it was possible that the veto function had complete independence from other neural activity (Libet 1999, pp.52-53). While the neural activity that occurs prior to the readiness potential is manifest only at an unconscious level, meaning we have no conscious control over the readiness potential's formation, the veto function allows for a limited form of conscious control over our actions that is independent of other brain states. Because we can consciously select the actions we choose to make via the veto function, we can be held responsible for our actions (Libet 1999, pp.54-55). We cannot, however, be held responsible for our thoughts, as the formation of thoughts occurs at a stage temporally prior to conscious awareness, and outside of the possibility of conscious control.

Rather than characterise Libet as endorsing hard determinism, it is better to think of his work as aiming to formulate a neural basis on which to support libertarianism, with this grounding his theory of responsibility. For Libet, it remains the individual agent who is ultimately responsible for her actions, as the veto function allows for a break within deterministic laws, with the conscious will having control over this break.

This is not, however, compatible with the libertarianism of Kane (Kane 1996). Agents must hold Ultimate Responsibility for at least some of their actions within Kane's libertarianism. This means that if an agent is responsible for an action, but there was another condition necessary for the action, the agent must also be responsible for that condition. It is in our character formation that Kane finds the origin of our free will. If for us to perform a particular action, we need to possess a particular character, we must be responsible for the formation of this character if we are to be responsible for the action. The idea of the "self-forming action" – an event during which our acts determine our characters – is all that is necessary; we need not be responsible for events prior to this, such as our own births, or the formation of the universe, thus preventing an infinite regress (Kane 2007, p.15). The conscious thoughts and unconscious brain states that lead to the actions we make are beyond our control, if we accept the truth of Libet's findings. Kane's libertarianism is thus vulnerable to the empirical challenges posed by Libet, though Libet himself offers a more modest version of libertarianism. In this way, Libet is working against the acceptance of hard determinism, rather than endorsing it.

Libet also challenges the claims of compatibilism from a neural perspective. As noted in Chapter 5, compatibilists claim that free will is not dependent on the falsity of determinism. Hume claimed that so long as a person is not in chains, she is free to choose a course of action in accordance with her will (Hume 1748, p.125). The truth of determinism would mean that she had no choice in the constitution of her will. However, rather than undermining free will, determinism is a necessary condition for free will. In an indeterministic world, Hume thought that all human action would be subject to chance rather than being controlled by the will (Hume 1748, pp.58-59). Max Velmans claims that even if all neural activity prior to action is unconsciously determined, we can still be deemed as acting in accord with our will (Velmans 1991), a claim which coheres with Hume's argument. Against Velmans, Libet thought the veto function must be consciously controlled in order for us to have free will (Libet 1999, pp.52-53). If all neural activity prior to action occurred unconsciously, then there could be no distinction between an action perceived as voluntary by an agent and an action which was the result of an epileptic seizure. Both actions would not be subject to conscious control and the agent would have no choice but to perform the action.

Stephen J. Morse argues that it is highly implausible to imagine that the brain works entirely deterministically, yet there is a function that is somehow independent and indeterministic (Morse 2008, pp.30-31). Work in neuroscience following Libet has revealed the deterministic nature of neural activity, meaning it is unlikely such an indeterministic function exists within the brain or mind. Patrick Haggard and Marcel Brass claim that the veto function is itself a deterministic process, existing within the dorsal fronto-median cortex (Brass & Haggard 2007, pp.9143-9144). Activity within this region inhibits certain actions. Thus, through a process of filtering in this region, we come to choose the actions we perform. The filtering process, however, is not subject to conscious control, nor is its activity somehow indeterministic. Brass and Haggard support Velmans' argument against Libet; the function is itself determined and does not offer hope for libertarian agency. Neither is the dorsal fronto-median cortex representative of consciousness. Activity here does not equate some kind of seat of consciousness from where decisions are made. The veto function is not, then, indeterministic or subject to conscious agential control, as Libet posited.

To conclude, Libet thought that while the readiness potential preceding conscious awareness was troubling for free will, the veto function meant that the conscious will was still responsible for controlling action, and this function was indeterministic. As Brass and Haggard have shown, however, it is highly unlikely that the neural processes equating the veto function are indeterministic. Instead, the dorsal fronto-median cortex, the activity of which works to inhibit some actions and allow for others, operates as deterministically as the rest of the brain. Libet's solution to the problem of libertarian free will fails when considering subsequent understandings of neural mechanisms.

7.3 Alternative Interpretations: Imaginings not Intentions

There are other ways of interpreting the data revealed by Libet's experiments. If we alter how we interpret the data, perhaps free will as formulated by Libet is salvageable. In this section, I examine alternative ways of thinking about the readiness potential and the veto function, and their relationships to the decision-making process. Mele's response to Libet is pertinent here. Mele's critique is within the bounds of Libet's own perspective; Mele is not introducing another viewpoint from which to argue against Libet but arguing against the assumptions Libet makes in his attempt to reveal how the data challenges traditional beliefs about free will. Essentially, both Mele and Libet are arguing for the same conception of free will, on which our conscious awareness somehow has causal functionality, though Libet argues from a libertarian perspective, while Mele is a compatibilist (though, as I argue, his position often slips between libertarianism and compatibilism). I argue that the alternative explanation offered by Mele fails to provide better grounds on which to argue for free will. This is partly as they are based on a misinterpretation of Libet's original claim, but also because these explanations do not show our conscious awareness to have some kind of independent causal function. Whether we think of the readiness potential as an imagining or an intention, we should accept its causal functionality.

If the readiness potential has causal functionality, then the origins of actions do not lie with the conscious will. Whatever constitutes our conscious will has been determined prior to our awareness of its constitution. If, following its formation, the conscious will does not have causal independence, then libertarianism is on shaky ground; there is no aspect of indeterminism within the neural activity behind human action. The libertarian would be left arguing that determinism may not be true at some other level, outside of neural activity, and that this allowed for our free will, despite us not holding the capacity to control this aspect of indeterminism.

Though Libet thought that the readiness potential revealed an intention to act that preceded conscious awareness of intention, it remains unclear why we should consider the readiness potential as representative of an 'intention'. There is no reason to assume that the brain activity constituting the readiness potential represents the decision itself (Morse 2008, p.30; Mele 2014, pp.12-13). It is perfectly possible, according to Morse and Mele, that the brain activity represents the brain's preparation to act, but that the decision itself is made later in the process. In order to understand whether the readiness potential necessarily led to the performance of an action, we would need to know whether the same brain activity occurred without any corresponding action. Libet did not consider this, and thus, for Mele, we cannot know at what point during the process we can consider the decision itself to have been made. This undermines the notion that the decision is made by the brain and then vetoed by the conscious will; we cannot know when the decision itself was made.

Mele also finds Libet's understanding of the veto function problematic (Mele 2014, pp.16-20). Libet came to posit the existence of the veto function through repeating the same experiment but asking subjects not to flex their wrists after deciding upon when to perform this action (Libet 1999, p.52). This is, for Mele, a fruitless task. Participants would never have intended to flex their wrists, as they knew they would never have to perform the action. The electroencephalography data obtained could show that the brain is thinking about or imagining an action, rather than revealing an intention to act, according to Mele (Mele 2014, p.19). In attempting to establish that agents can consciously veto an impulse to act, all Libet is revealing is that subjects will imagine a particular action without having any intention of ever performing the said action.

For Mele, as for John R. Searle, a distinction should be made between general and proximal intentions (Searle 2000, pp.17-19; Mele 2004, pp.19-23). General intentions are those in which a person understands that at some point they will probably perform an action, but do not know when this will occur. Proximal intentions are those in which a person has decided to act now. A subject in one of Libet's experiments would have a general intention to act throughout the experiment. However, the presence of proximal intentions in the experiment is more complex. Mele writes that in go-signal reaction time tests – tests in which participants are given a go signal such as a tone sounding and must respond immediately by doing something such as pressing a button – the time between the signal and the muscle burst is less than 231 milliseconds (Haggard & Magno 1999; Mele 2014, p.21). For Mele, this means that time between a proximal intention and an action can be much shorter than Libet suggests, who saw the readiness potential occurring on average 550 milliseconds prior to action. Furthermore, the process is not necessarily occurring at an unconscious level; subjects are consciously aware of the whole process in go-signal reaction time tests.

Likewise, Dennett writes that "what Libet discovered was not that consciousness lags ominously behind unconscious decision, but that conscious decision-making takes time" (Dennett 2003, p.239). Not all actions are the same, and the decision-making processes that characterise certain actions do not appertain to all other actions. Dennett uses the example of a tennis player returning service. A tennis ball being served by Venus Williams can cross from baseline to baseline in 450 milliseconds, 50 milliseconds less than the average time of the readiness potential's initiation (Dennett 2003, p.238). To return this service, Venus Williams' opponent must be able to visually process the situation and then prepare her body for action in less than 450 milliseconds. That this is possible is due to the way in which the conscious decisions have been made, according to Dennett. The tennis player consciously commits to a course of action beforehand, and then during the game of tennis, allows "reflexes" to follow through with this course of action (Dennett 2003, p.238), in much the same way Searle and Mele describe proximal intentions (Searle 2000, pp.17-19; Mele 2004, pp.19-23). These are not the reflexes explained in the previous chapter, which are involuntary responses to stimuli, but reflexes that have been consciously predetermined. The speed in which humans are capable of executing such actions allow for our capacity to play ensemble music or respond to others in conversation (Dennett 2003, p.239). What Libet uncovered, Dennett posits, was that when actions are considered consciously, the whole process takes longer. When we think about a specific movement our wrists are making, the process will take much longer than when we are attempting to return service while playing tennis. In actions requiring us to respond quickly such as in a tennis game, it is unlikely we would be consciously considering the placement of our wrists at all.

Mele thinks that there are two possibilities within Libet's experiment regarding proximal intentions (Mele 2014, pp.22-23). It could be that proximal intentions are not involved, and the subject uses the general intention to act in conjunction with the urge to act at a particular time, bypassing the conscious proximal intention. In this case there is no unconscious intention, only a conscious general intention leading to an act at a certain time. Alternatively, it is possible that proximal intentions are involved, and the 231 milliseconds noted in go-signal reaction time tests is close to the point at which subjects in Libet's experiments become consciously aware of the intention to act, 200 milliseconds prior to acting. In either scenario, Mele claims that conscious decision-making would be actively involved in the process. The conscious will would still be the cause of our actions.

As Mele admits, his work is speculative (Mele 2014, p.23), as was Libet's. The claims he is making are based on what he considers to be a more reasonable interpretation of the empirical data. Mele states that the evidence better supports his claims than it does Libet's. Mele attempts to show that the conclusions drawn by Libet are misguided. We cannot know at what point during the process the decision is made, according to Mele. Brain activity prior to conscious awareness of action could represent imagining of the act, or just a preparation to act, rather than the decision itself. There is, however, nothing in this that contradicts Libet's position. Libet thought that the decision itself was made at the point at which a person became consciously aware of her intention to act, and chose to either veto or perform the act.

A problem with Mele's account is that he slightly mischaracterises Libet's argument. Mele writes that Libet's challenge to free will rests on the claim that "we make all our decisions unconsciously" (Mele 2014 p.21). Others have used Libet's work to support this claim (Wegner 2002; Harris 2012), but Libet did not make it himself. While the origin of an intention to act is indeed formed unconsciously, by the time a person comes to act, consciousness is involved in the process. Libet thought that conscious awareness did allow for free will via the veto function. Thus, we are able to consciously decide which actions we choose to make. Moreover, the claim that the readiness potential represents a decision is also mistaken. Libet never argued that the readiness potential was the point at which a decision is made; the decision itself is made when the individual becomes consciously aware of the urge. Mele's claim that the readiness potential is an imagining of the act rather than the decision to act is perfectly compatible with Libet's picture of events.

This alternative explanation of events does nothing to advance on Libet's own reading of the data. First, for the most part, Mele does not contradict any of Libet's claims; Mele's line of argument is congruent with Libet's original argument. Second, Mele's interpretation of the data tells us little about free will. The most significant finding of Libet's experiments is that our intentions appear to form in our brains before we become consciously aware of our intentions. Mele does not deny that this neural activity occurs prior to conscious awareness. Arguing that the neural activity represents an imagining rather than an intention or decision, however, does not counter the basic premise of Libet's argument. If we are to argue that the readiness potential is only an imagining of an action, and that the decision itself is made separately, then we must show that the decision somehow has causal independence from the imagining. The alternative picture of events drawn by Mele does not show this. Rather, Mele shows that we have certain conscious desires leading to certain actions. As work in neuroscience following Libet has shown, however, these conscious desires are determined by prior neural activity (Brass & Haggard 2007; Roth 2010). The conscious desire has no causal independence from this activity, whether we think of the activity as representative of an imagining or an intention. Libet and Mele are on the same side of the argument here, though they may not realise this. Neither of their interpretations of the data, however, succeed in showing that our decision-making is independent of other neural processes.

Elsewhere, Mele attempts to show how the conscious will is causally functional, as I discuss in 7.4. Here, however, through arguing that the readiness potential can be considered as an imagining, Mele is not showing the will to be causally functional. The original problematic finding from Libet's experiment – that neural activity precedes and determines the conscious will, meaning the conscious will has no causal function – remains unresolved. Mele's alternative picture of events gives us no reason to think that consciousness' functionality has been restored. Moreover, there is no reason to suppose that this picture better supports an argument for free will, if we are to accept that conscious control is a necessary condition for free will. Mele does nothing to show that the readiness potential is somehow consciously initiated. Thus, neither Mele nor Libet, in their respective readings of events, offer us reasons to accept that the conscious will is causally functional, and that it is this function that enables us to have free will. The conscious will does not have causal functionality, independent of other causal factors.

In summary, there are alternative ways of interpreting the data obtained from Libet's experiments. It could be that the readiness potential is representative of an imagining of the act rather than an urge to act. This picture of events is as realistic as Libet's, according to Mele. However, there are two flaws here. The first is in arguing against the claim that the conscious will is functionless. This was never Libet's claim. Libet thought that the veto function restored a functional role for the conscious will. The second flaw is in arguing against Libet's assertion that the readiness potential is the decision itself. This was also not Libet's argument. The decision itself comes after the readiness potential, when the agent becomes consciously aware of her intention to act. Beyond these inaccuracies in Mele's response to Libet, however, there is little reason to think that the alternative interpretation allows for the truth of free will, as consciousness somehow has causal functionality. For these reasons, the alternative explanations fail to provide a better interpretation of the data, and thus do not provide better grounds on which to argue for free will.

7.4 Motivations and the Reasons for Action

Another way of responding to Libet's challenge to free will is to reconsider the internal motives of agents. I examine here a different aspect of Mele's argument, in which he further attempts to show how the conscious will is functional. According to Mele, Libet's work does not take into account the importance of motivation (Mele 2014, pp.13-15). I examine the various motivations that subjects in Libet's experiments may or may not hold, and assess how these motivations may influence the results of the experiment. Ultimately, I argue that there is no reason to assume that motivational factors are not also determined outside of conscious awareness, and that Mele fails to re-establish the conscious will as holding causal functionality.

Subjects within Libet's experiments can have little motivation to flex their wrists, they only do so in order to comply with the guidelines of the experiment. This is problematic for the psychologists Julius Kuhl and Sander L. Koole who argue that the subjects are merely responding to the will of those in charge of the experiment, rather than revealing anything about their own wills (Kuhl & Koole 2004, pp.419-420). At no point during the process would it be more appropriate to flex the wrist than at any other, meaning that the time at which the wrist is flexed will be chosen randomly. Mele compares this to choosing a jar of nuts in a shop (Mele 2014, pp.12-13). If all the jars are identical, there can be no reason to choose one over another. The decision of which one to take will thus be made randomly, and at an unconscious level. We would most likely tell someone who asked why we chose a particular jar that we did not know, due to the unconscious way in which it was chosen. Some decisions may be made unconsciously, but Mele argues that it is wrong to infer that *all* decisions are made unconsciously. If there are reasons for us to prefer to perform certain actions and not others, Mele thinks it likely that the decision-making process would involve a greater level of conscious deliberation. Without a reason to perform an action, there is no need for conscious deliberation. As there was no reason for subjects to flex their wrists in Libet's experiment, the experiment only reveals the unconscious thought processes occurring behind actions requiring little conscious deliberation. In neuroscience, Roth also asserts that Libet's results are only pertinent in regard to "short-range and pre-programmed movements" (Roth 2010, p.239). Deliberation over an action would likely affect the results, as conscious thoughts influence the rising of the readiness potential, as Mele argued.

Through the introduction of motivating factors, Mele is attempting to show how the conscious will can hold a causal function. Mele and Libet are arguing for the truth of free will on the same basis. Both agree that in order for us to have free will, we have to be capable of consciously choosing our actions, and that if all of our actions are unconsciously determined, we cannot have free will. Mele, however, is not a libertarian but a compatibilist. For Mele, the truth of free will is dependent only on our ability to make rationally, informed decisions about our courses of action (Mele 2014, p.78).

Mele's argument is extended through the example of selecting seats on an airline (Mele 2014, p.52-53). As he has a conscious preference for extra legroom, Mele posits that he then consciously chooses the exit row seat. Consciousness is thus posited as a functional stage in the process of causation. However, there is no reason to assume that this conscious preference was not also unconsciously formed. Again, Mele attempts to argue with the incompatibilist position on incompatibilist grounds; rather than argue against Libet from a compatibilist position, Mele accepts Libet's basic premise that consciousness must be independent of previous neural activity if we are to have free will. A Humean argument could be introduced here (this is explored further in 7.7); determinism is a necessary condition for free will, as without it, our actions would be subject only to chance. The deterministic and unconscious origins of the will do not therefore undermine free will, but rather reveal the deterministic relation between the will and human action. So long as there is coherence between the will – once it is made conscious – and action, the will can be considered free. Mele, however, does not make any such argument, but attempts to show that the conscious will can be the cause of human action.

There is a problem with attempting to re-establish consciousness as a cause of action. Though Mele argues the conscious will does not have to be some kind of magical entity (Mele 2014, pp.85-86), existing apart from the rest of the self, in positing consciousness as a cause of action, Mele is reverting back to the idea that consciousness is a *causa sui*, independent of other neural processes. If this is not the case, then Mele must concede that the conscious will has been shaped by prior neural processes, and we are left with the original problem posed by Libet. There is little evidence to suggest that any kind of conscious intention works independently of other neural activity and somehow influences subsequent neural processes. Though Roth argues that results would change if we measured actions requiring more pre-planning than the movements in Libet's experiments – as deliberation would affect the ordering between conscious awareness and the readiness potential, with there being interaction between the two rather than a linear process – he also argues that processes of deliberation are outside of conscious control (Roth 2010, p.239). The conscious will has no choice over the desires and emotions – generated largely unconsciously – influencing the deliberative process. In arguing that the conscious will can be considered a cause, Mele is falling back on a libertarian conception of free will, rather than making the compatibilist case.

To conclude, another way of interpreting Libet's data is to look at the internal motivational states that influence human action. Human action is influenced by the motivating reasons that undergird it; in the absence of motivation, actions would be performed more or less randomly. This is the case in Libet's experiments, according to Mele. During Libet's experiment, it is never more appropriate to flex your wrist at one moment rather than another. Hence the decision of when to flex will be made at random. If we consider human action within a context in which motivating reasons are at play, then action comes to look radically different. In making this case, however, Mele is once again attempting to establish consciousness as a cause of action. There is little reason to suspect that motivating reasons offer some kind of independent form of causation. At the internal level, motivations will most likely have neural causes, and at the external level, the reasons that motivate an agent are independent of the agent. Neither offers better ground on which to argue for free will.

7.5 The Absence of the Cartesian Theatre

As stated previously, Libet's work only affects a certain conception of free will. This conception is both dualistic and libertarian; dualistic as it distinguishes the mind from the body, imagining that the mind as a separate entity must control the body, and libertarian as this entity must be independent of causal determinism. Daniel Dennett calls this entity the "Cartesian theatre". The Cartesian theatre – named as such due to Descartes' dualistic understanding of the mind and body – is a seat of consciousness from where human action can be controlled. Perhaps, however, the mind is not a separate entity and there is no external point from which the mind controls the body. Some such as Dennett argue that there is no Cartesian theatre and, furthermore, that no such entity is necessary for us to have free will. In this section, I assess arguments for free will that do not depend on the existence of a Cartesian theatre, and argue that while libertarianism requires the existence of such a centre of consciousness, compatibilism does not.

For Dennett, there is no seat of consciousness from where commands are issued (Dennett 2003, pp.232-236). Whereas Libet thought that our conscious awareness represented a significant temporal stage in the process between intention and action, Dennett sees consciousness as being spread out across the entirety of the process, not contained within specific temporal moments. There is not some kind of self existing outside of the rest of the process, watching over it as though it were a spectator in a theatre, and who acts as the ultimate arbiter of decision-making. The individual is rather included in each part of the process. In the example of a subject in Libet's experiment, the individual is composed of the eyes, wrist, and brain, along with all of the interconnecting parts.

For Libet's work to have any bearing on the question of free will, we must accept the premise that the actual 'you' exists in some part of the brain. This 'you' is ultimately responsible for all decisions. Dennett suggests three possible places for this 'you' to exist: the faculty of practical reasoning, the vision centre, and the Cartesian theatre (Dennett 2003, pp.232-236).[1] If 'you' are in any one of these places, there will be a delay in the activity of the other regions being processed and the information sent to 'you'. In the faculty of practical reasoning, 'you' would have to wait for the activity of the vision centre to be completed, and vice versa if 'you' are in the vision centre. Alternatively, 'you' could be in the Cartesian theatre, watching over the whole process.

Dennett argues, however, that this Cartesian theatre does not exist; there is no place within the brain or mind that represents such a command centre for conscious thought, either spatially or temporally. Consciousness is instead "broken up and distributed in space *and time* in the brain" (Dennett 2003, p.238). This formulation of consciousness is largely in agreement with Roth's interpretation of the empirical data. Roth argues that the traditional view of the cerebral cortex as a command centre from which all directives are issued is mistaken (Roth 2003b, p.115). For Dennett, we are not outside of the loop, our being is constituted by each part of the loop, and both our free will and moral agency are spread across the entirety of the process (Dennett 2003, p.242). Free will, then, cannot be measured in instants of time, as in Libet's experiment; for Dennett, it needs to be understood as a phenomenon that exists and develops across time.

These arguments are consistent with Dennett's compatibilism, as explicated in *Elbow Room* (1984). Freedom, for Dennett, has an evolutionary aspect; it is something that develops as we learn from experiences. Whether or not determinism is true, a human

[1] It should be noted that Libet did not think of his theory of mind as being Cartesian (Libet 2006, p.324). For Libet, all neural functions are material; he did not posit that the veto function was an immaterial object existing outside of the brain itself.

action has consequences that extend beyond the act itself. Thus, determinism should not deter us from considering the importance of our actions or lead to fatalism; we should not resign ourselves to our fates if we consider determinism to be true (Dennett 1984, pp.102-107). An agent in a deterministic universe is still able to hold desires and to work to realise these desires. Without acting on plans, the agent could not realise any desires. Dennett proceeds to endorse Frankfurt's arguments for compatibilism, arguing that we do not need alternative possibilities for our actions to be a result of free will (Dennett 1984, p.132). Providing that there is coherence between intentions and actions, a person is acting according to her free will. In any one temporal instant, there need not be alternative possibilities. However, we must learn from experience; from our failures we can learn methods for improving our behaviour (Dennett 1984, pp.142-143). Though there need not be alternative possibilities in one given moment, the potential of future alternative possibilities means that we are free to act in preparation for these alternatives. This allows for us to possess free will even if determinism is true.[2]

We do not, as Libet tacitly implied, need a break within deterministic laws, which the conscious will has control over. Such a break is not possible – as there is no Cartesian theatre from which the break could be controlled – nor is it necessary. If an agent is able to act in accord with her will, and, learning from experience, to act differently in the future, this is all that is necessary for her free will. A conscious will sitting apart from the process, yet also controlling it, is not what constitutes free will for Dennett. Rather than our bodies being controlled by a conscious will located within some kind of command centre, our will is instead spread across the body. This is an interpretation of the empirical data in agreement with Roth's assertion that the cerebral cortex is not the command centre of the human brain (Roth 2003b, p.115), and that functions allowing for decision-making are distributed across various regions of the brain.

What was described by Libet as the readiness potential is thought by Roth to be preceded by activity in various regions of the motor cortex (Roth 2010, p.234). The basic process between the brain and action leads from the motor cortex to the spinal cord, and from the spinal cord to the muscle necessary to move the body part. "Willed" actions, however, involve activity in further neural regions. The basal ganglia are of importance here, as they store memories of previous successful actions (Roth 2010, pp.235-236). Alternative

[2] However, as I argue later, this does not grant us the level of control over our thought processes that Rawls requires for us to be rationally autonomous.

courses of action are inhibited by the basal ganglia so that the course of action most likely to succeed is selected. Inhibition of alternatives is enabled by the release of dopamine. It is the limbic system that has control over the release of dopamine, particularly the amygdala with its storage of emotion (Roth 2010, p.236). These emotional memories inform the striatum – part of the basal ganglia – which then releases dopamine to allow for inhibition of some options and the excitation of another. Only after this does the readiness potential arise. A system of final checks is made between all of the aforementioned neural regions before the signal to act is sent from the motor cortex to the spine (Roth 2010, pp.236-237). Within this process, Roth does not identify a centre of consciousness either watching over all the events or waiting to select a course of action via the veto function. Whereas the cerebral cortex was once thought of as such a centre of conscious thought, Roth disputes this; instead, the processes occurring across the brain relate to our experience of conscious awareness (Roth 2003b, p.115).

However, Dennett does not resolve the problem of the origins of our intentions being beyond our conscious awareness. This is not to say that this necessarily poses a problem for Dennett's compatibilist position, though it does challenge the libertarian perspective on free will. Libet claimed that while we are responsible for our actions, we are not responsible for our thoughts. Whether we consider this a problem though depends partly on the conception of free will we endorse. Compatibilists may not view this as a problem. This will be explored further in 7.7.

There is, for Dennett, no Cartesian theatre from where decisions are made independently of other neural processes. This conception of the brain and its workings is consistent with the empirical evidence. Dennett's argument here is strengthened through its reliance on compatibilism, rather than slipping between libertarianism and compatibilism, as Mele does. Arguments for libertarian free will are difficult to reconcile with the empirical evidence at the neural level.

Helen Steward, however, has argued for a libertarian conception of agency in which no Cartesian theatre is involved. Steward argues that a conscious decision does not need to precede an action on her libertarian view of agency (Steward 2012, pp.43-49). Bodily movements are subsequent to actions; our body moves because we act, we do not act because our body moves. For an action that occurs over a certain temporal period to be considered free, there needs to be alternative possibilities throughout the process. Removing alternative possibilities at any point during the process would mean that the action lost its freedom. Framed in this way, the agent is “constantly settling what happens from one moment to the next” (Steward 2012, p.46). Even if the power to do so is not utilised, it is important that the agent is constantly able to alter the action. Libet’s findings do not challenge this conception of agency, Steward argues, because on her view, actions must be free from beginning to end, rather than being initiated by a conscious will that triggers a deterministic course of action (Steward 2012, pp.46-47). Actions are, then, processes rather than events. The whole process must be free, and the conscious will’s involvement is simply one part of the process. For Steward, Libet’s work does not challenge libertarianism.

The evidence within neuroscience, however, suggests that the process is deterministic. Though Steward may be right to argue her conception of agency is compatible with Libet’s findings, it coheres less with the subsequent work of Haggard and Roth. In these works, neural processes are understood as working deterministically, whether or not determinism at the universal level is true. Upon the evidence offered by Haggard and Roth, it is not that agents are able to settle what happens from one moment to the next, and that this is somehow indeterministic. The conscious will may not trigger a deterministic course of action, but this does nothing to support Steward’s view. Rather, neural processes are occurring in deterministic circuits. Within the empirical data, there is nothing suggesting that neural processes are indeterminate in the way that Steward requires.

From examining the empirical data offered by Roth, it can be ascertained that there is no Cartesian theatre; the cerebral cortex is not the centre of conscious awareness, as once thought, but various neural regions are linked with our experience of consciousness. Whether free will can exist without the Cartesian theatre depends on our formulation of free will. Certainly, the libertarianism of Kane and Libet is dependent on some kind of Cartesian theatre. Steward’s formulation of libertarianism, however, does not require a Cartesian theatre. While Steward’s libertarianism is unthreatened by Libet’s result, it is more vulnerable to subsequent findings in neuroscience. There are deterministic physical processes occurring in the brain. None of these processes are subject to the kind of indeterminism that Steward requires. Dennett shows, however, that compatibilism is not dependent on the existence of a Cartesian theatre. All that is necessary for free will, for Dennett, is that we act on our own volitions, and through learning about the effectiveness of our actions, open up new possibilities in the future. While I argue this conception of free will coheres with the empirical data, it does not resolve the central problem revealed through the data: we lack control over our thoughts. This is explored later in this chapter.

In summary, Dennett's response to Libet's work can be explained through two key arguments. Firstly, human actions are not uniform and monolithic, there are vast variations between different actions, and certain actions require differing levels of conscious involvement. All Libet discovered, according to Dennett, is that actions requiring a good degree of conscious thought require a longer process in which the decision to act can be made. Secondly, there is no Cartesian theatre which we inhabit. Our consciousness is spread out across the human brain, and so we cannot look to a particular part of the brain in order to assess whether or not we have free will. Dennett's claims are coherent with the empirical data; Roth postulates that there is no command centre within the human brain. Instead, there are networks concerned with voluntary actions, and others relating to involuntary actions, with each part of a network being equally involved in the process. Only under compatibilism can free will be considered independent of a Cartesian theatre.

7.6 Social Influences on the Constitution of the Will

One way of responding to the challenge posed to free will by neuroscience is to look at the ways in which social influences inform the will. Some argue that awareness of environment, which develops across time through learning from experience, is what leads us to develop free will. Mele and Dennett, along with Cacioppo and Berntson, have all made this argument though in different ways. Learning from our experiences, we open up possibilities of acting differently in the future. This is an important aspect of the compatibilism of Mele and Dennett. Alternatively, it can be argued that the strength of social influences further undermines the ability of an agent to act according to her own will. I begin by assessing claims made in psychology regarding social influences on our intentions, before assessing claims made by Mele and Dennett. I claim that, from a neural perspective, social factors cannot be viewed as favourable to arguments for free will.

In psychology, Julius Kuhl and Sander L. Koole have criticised the setup of Libet's experiments (Kuhl & Koole 2004, pp.419-420). The command given to the participants – to flick their wrists at a self-chosen time – is already a limit to their personal freedom. Some participants are liable to mistake external commands for the internal will. Kuhl and

Koole cite other experiments in which Kuhl was involved to support this claim (Baumann & Kuhl 2003; Kazén et al 2003). Some people, termed here as "state-oriented", suppress their own will in response to authoritative commands. As Libet's experiments do not account for the personalities of participants, it is difficult to know to what extent participants saw the action performed as resulting from their own will; it may be that participants saw the action as deriving from the will of those issuing the command. Here, the focus of the argument shifts from constraints on the will to the origins of the will.

Libet himself thought that the will had mysterious origins (Libet 1999, p.54). While we could be held responsible for our actions, due to our conscious involvement in our choice of actions due to the veto function, the origin of the neural processes from which the will derives is difficult to locate. Social neuroscientists trace our intentions back to our social world. Cacioppo and Berntson argue our will is dependent on our sociability (Cacioppo & Berntson 2012). While consciousness in a singular brain is epiphenomenal, when brains are socialised, consciousness becomes functional. Through interrupting the neural patterns of others, we come to change their behaviour. The will, as expressed through our consciousness, has then a social function. In isolation, we would therefore not have free will, but on entering a social world, our consciousness becomes functional, further enabling the development of free will. This view, however, is entirely speculative. There is no empirical evidence to support a thesis on the origins of the will.

Mele also views social influences as not undermining arguments for free will. Experiments such as the Stanford prison experiment and Stanley Milgram's studies suggest that humans will inflict cruelty on others merely because they are told to, or because the environmental conditions are such that they feel compelled to (Mele 2014, pp.52-76). This does not deny the free will of the participants, according to Mele. It is true that, for the majority of time, individuals behave according to the expectations of a situation. Thus, Mele writes that during the attacks on the World Trade Centre in 2001, the passengers complied with the hijackers because compliance is expected of passengers on a flight (Mele 2014, pp.62-63). However, Mele claims that these attacks could not be repeated. Once people are aware of possible outcomes, their behaviour will change. With the knowledge of what happened during those attacks, passengers would no longer comply with the hijackers. This is shown partly, Mele argues, by the behaviour of passengers on United Airlines flight 93, who heard about the prior attacks, and attempted to regain control from the hijackers (Mele 2014, p.65). Education and awareness of an environment allow for the modification of behaviour, and allow for our free will, rather than deny it.

Mele tacitly concedes, however, that the environmental circumstances did dictate the wills of the victims on the planes during the World Trade Centre attacks. They were accustomed to an environment in which they were expected to obey the orders of the flight crew, and in the absence of these orders, they instead obeyed the orders of the hijackers. Their individual wills were thus environmentally determined. Future passengers will have an awareness of these events, and will therefore disobey hijackers, according to Mele. Even if we accept the premise of this argument – that the passengers did not rebel against the hijackers due to their being accustomed to obedience, rather than, for example, their extreme fear of the terrorists – there is no reason to assume the wills of the imagined future passengers overthrowing their hijackers are not also environmentally formed. Their awareness of previous experiences has led to the decision made. Without this awareness, they would not be able to reach this decision. In both cases, the wills of the agents are subject to the environment, and their awareness of previous experience. Mele's argument here is congruent with those of Cacioppo and Berntson, yet although social relations can affect and alter human behaviour, it is not the individual will that is altering the social environment but the social environment changing the individual will. Incorporating social influences does not, then, advance Mele's argument for free will.

Dennett also argues that social and environmental factors enhance rather than threaten free will. Alternative possibilities in temporal instants are not necessary conditions for free will for Dennett. Instead, what is necessary is our ability to learn from experiences, opening up alternative possibilities in the future (Dennett 1984, pp.142-143). According to these arguments, however, the first person to experience an event has less freedom than the second. If particular knowledge is required to survive both event A and the subsequent event B, but the only way for this knowledge to be gained is through the occurrence of the event, those involved in event A have no chance of survival. While those in event B can exercise their free will through the knowledge obtained from event A, those in event A have no such free will. On this formulation of free will, those existing temporally prior to others yet experiencing the same situation have no free will. However, subsequent participants in events could never be sure that the rules of the game are the same. Though they may have more options available to them due to knowledge obtained from past experiences, they could not be certain of their success. It is then difficult to state at what point people are acting according to their free will, and when they are acting under conditions through which their free will is negated.

To sum up, emphasising social influences does not bolster arguments for free will. Dennett and Mele both argue that, for the truth of free will, alternative possibilities are not necessary in single moments. Instead, because we can learn from our experiences, we can allow for alternative possibilities in the future. However, there is no single point at which we can, then, say that a person is definitely acting according to her free will. The situation may not be exactly the same as previous situations. We can thus not know at one point we consider someone to be acting according to her free will.

7.7 Two Compatibilist Defences

As has been seen, the empirical evidence challenges the libertarian conception of free will. Kane's libertarianism, which demands Ultimate Responsibility, is undermined by the empirical evidence, as there is no uncaused cause within the brain; there is no point at which an agent could be said to have performed a self-forming action, as her actions were determined by prior neural activity. Less demanding conceptions of libertarianism, such as posited by Libet, are also challenged; subsequent research in neuroscience has found that the veto function itself is a deterministic process. Thus far, there have been few reasons presented for re-evaluating the claims of compatibilism in response to neuroscience, other than the rejection of arguments for the importance of social influences in 7.6 (though these arguments cohere with both libertarianism and compatibilism). In this section, I examine two compatibilist positions, those of Hume and Frankfurt. I argue that Hume's position on free will tells us only whether an act can be considered free at the external level. This position on freedom does not speak to the internal neural aspects of the will, and whether we can speak of this will as being free. Regarding Frankfurt, I argue that while it is plausible to argue for responsibility on the basis of our first and second-order desires cohering, this position does not resolve the issue that the origins of our thoughts are prior to conscious awareness. Voluntariness, rather than freedom, I argue, is ultimately what Hume and Frankfurt are arguing for. Whether or not Frankfurt's argument for responsibility holds, the implications of neuroscience remain problematic. While our actions might be voluntary, we hold little control over the thought processes behind them.

At first sight, it does not appear that the claims of compatibilism are threatened by neural evidence. Providing that an agent's intentions align with her actions, the compatibilist can argue that her actions are free. Hume went further; not only was free will compatible with determinism, determinism was a necessary condition of our free will. I begin by addressing this claim and argue that it tells us only about the level of freedom in an external situation.

For Hume, an indeterministic world would be one in which the individual had no control over her actions, as everything would be subject to chance. Because of this, there could be no necessary connection between a person's moral character and her actions (Hume 1748, p.65). Determinism means that there can be complete uniformity between cause and effect; if the conditions pertaining to one cause are precisely the same as another, the same effect will be produced in both cases. Individuals could thus know the results of their behaviour, and there would be a necessary connection between a person's character and her behaviour. Without this connection between character and behaviour, human action would be subject to chance rather than the individual will. A universe in which all human actions – along with all other physical events – were the result of randomness would not allow for free will. Determinism is, then, a prerequisite for free will, as in a deterministic universe, humans can control their behaviour with a level of certainty. In a deterministic universe, the will can determine the action.

Hume goes on to say that all that is required for an individual to be free is her unconstrained power to choose (Hume 1748, p.69). An individual may choose to rest or to move. In a deterministic world, we can be sure that when we choose to move, our body will perform the movement chosen, as there is uniformity between cause and effect. The only cases in which a person's actions can be considered unfree, Hume says, are when a person is physically constrained, such as a prisoner in chains (Hume 1748, p.69). A person not restricted by chains is able to determine her own actions through her ability to choose. The absence of constraints allows for a necessary connection between the will and action. This is what defines freedom of the will.

Suppose Hume was to respond to the concerns raised by Libet's experiment. Hume might say that the flick of the wrist was free because nothing prevented the participant from performing this action. The timing of action in relation to neural processes and conscious awareness is irrelevant. When the participant flicks his or her wrist, the necessary connection between the will and the action is revealed, and hence the action can be considered free.

There is both an internal and an external aspect to Hume's view of freedom. The internal aspect is the relation between the will and the act. Determination of the act by the will is needed for an act to be an act of free will. The external aspect is the lack of impediments to an act's performance. If nothing is externally preventing an act being performed, then it is an act of free will. If something external compels the act, then it is not an act of free will. On Hume's view, the formation of the intention is unimportant. It is not in the formation of our will that its freedom lies, but in its realisation, if external conditions allow for this realisation. Libet thought that if all intentions to act were formed unconsciously, however, there could be no way of distinguishing between actions performed as the result of an epileptic fit, and actions performed as a result of agential volition. Both would need to be considered free.

Consider the following three actions:

A. An action I desired to perform.
B. An action I desired to perform but would have preferred to perform another action.
C. An action I did not desire to perform but was not compelled to perform due to external conditions (for instance, actions resulting from an epileptic seizure or delirium).

Action A can clearly be considered free on Hume's view. Action C does meet Hume's criteria. For an act to be an act of free will, it should reveal the necessary connection between the will and the action. Action C does not reveal this connection; we should, therefore, not consider it an act of free will.

Action B proves more difficult, however. Imagine a prisoner who desires to read a book on the other side of her cell but would much prefer to leave her cell and go home. The prisoner in chains can do neither. We can say, then, that she does not act of her free will. The unchained prisoner, however, can go and read the book. On Hume's view, we could say that the prisoner acts of her free will, as nothing was preventing her from reading the book. However, the act does not reveal her overriding will to leave the prison. The unchained prisoner's everyday acts in her cell never reveal this will. Perhaps, then, we should not consider either prisoner as acting of her free will. On this view, Action B is not an act of free will, as it does not reveal the necessary connection between the will and the act.

This view leads to incoherence, however. If we perform an action, but would have preferred to perform another action, it can be argued we did not act according to our free will. This captures many actions across our daily lives. Employees going to work who would prefer to spend the day at home would be seen to act against their will. Such people may not be physically restricted from choosing the alternative course of action, but the consequences of performing the action may be so severe as to act as a physical restraint (being unable to feed or house their children, for instance). Dropping the condition that there must be a necessary connection between the will and the action forces us to say that Action C is an act of free will, however. If we drop this condition, we stretch the definition of freedom to be so broad that it becomes meaningless. If we keep the condition, freedom becomes so narrow that it captures very few of the actions we make in life.

Considering any externally unrestricted act as being free, we are left unable to respond to complexities relating to the constitution of the internal will. This is not the case with Frankfurt's compatibilism. Here, inner motivational states are of more importance than on Hume's view. For Frankfurt, it is not the availability of alternate possibilities that allows for our free will, but the human capacity to hold "second-order desires" (Frankfurt 1971, pp.6-7). Generally speaking, humans do not merely act on instinct; rather, they hold certain immediate desires (first-order desires) while simultaneously holding a desire to hold – or not to hold – this desire (second-order desires). Someone may want to eat a slice of cake, whilst also acknowledging that this desire conflicts with dietary plans. The agent would then hold a desire not to desire to eat the cake. Our ability to hold such desires allows for our free will, according to Frankfurt, as it enables us to subject our decisions to rational deliberation (Frankfurt 1971, p.14). An agent who did not hold second-order desires would not be capable of acting according to her free will. Frankfurt describes such an agent as a "wanton"; an agent who acts only on first-order desires (1971, p.11).

Alternate possibilities are therefore not a necessary condition for our free will, for Frankfurt. Providing we hold both first-order and second-order desires, and these desires cohere, we are free. In Frankfurt's essay 'Alternate Possibilities and Moral

Responsibility', he establishes why alternate possibilities are not needed through the use of four examples (Frankfurt 1969). In the fourth example, Jones has decided on a course of action, which Black is determined Jones will perform (Frankfurt 1969, pp.835-839). Black decides that, should Jones decide against performing the action, he will intervene to force Jones to commit to the action, by hypnosis or a magical spell. Coming to the time of the event, Jones does not stray from the course of action, meaning Black need not intervene. Frankfurt's argument here is that we can still hold Jones responsible for his action. Though Jones had no alternative to act other than he did, his moral responsibility is not negated. It can be said, then, that alternative possibilities are not a necessary condition of moral responsibility. This is due to Frankfurt's belief that free will – on which moral responsibility depends – is based on this relationship between desires, rather than the undetermined will. As there is consistency between the first-order and second-order desires, Jones is acting of his free will, and can therefore hold moral responsibility.

Earlier in the essay, Frankfurt uses another three examples (1969, pp.831-833). In these examples, Jones intends to perform a task, and is then threatened into performing said task, which he proceeds to do anyway. Jones performs the task in the first example because he has an unreasonable character; once he has decided on a course of action, he will stick to it regardless. Here, Frankfurt argues Jones still holds responsibility for the action, as his prior intention still holds. In the second example, Jones has a timid character and is profoundly affected by the threat. He therefore commits the action because of the threat; his earlier intention to commit the action is no longer the cause of his action. Frankfurt argues, then, that Jones does not hold moral responsibility in this case. In the third example, Jones has a reasonable character and, though he is affected by the threat, he continues to deliberate over whether to carry out the action. Jones decides that he will on the basis of his original intention rather than the threat, and therefore holds a degree of responsibility, according to Frankfurt.

In examples one, three, and four, Jones can be said to hold moral responsibility as his first-order desires are in accord with his second-order desires. His desire to perform the action is a desire he desires to hold. In the second example, Jones does not hold moral responsibility, as he no longer holds a desire to perform the action; he performs it only out of fear of the threat. Nothing is undermining the freedom of the will if both the desire and action are determined, meaning alternate desires and actions are unavailable. Frankfurt acknowledges that freedom comes easier to some than it does to others, and that those beings – whether human or another kind of being – incapable of forming second-order desires cannot be considered to have free will (Frankfurt 1971, pp.14-17). In this sense, we have no choice over whether we are free: some possess free will, some do not, while others struggle to achieve it.

While Frankfurt is correct to view the alignment of second-order and first-order desires as allowing for responsibility, it is not necessarily constitutive of freedom. To argue that it is constitutive of freedom forces us to accept a particular formulation of freedom. Rather than consider this as freedom, we should, I argue, think of it as voluntariness. There are two reasons for making this argument. First, Serena Olsaretti argues that freedom and voluntariness are two different concepts, and that one does not imply the other (Olsaretti 1998, p.53). Freedom refers to the choices with which we are faced, while voluntariness relates to the choices we make. In what follows, I do not endorse Olsaretti's understanding of voluntariness, but acknowledge the distinction between the two is important. Second, John Martin Fischer argues for a position he calls "semicompatibilism". This is the view that while compatibilism – following Frankfurt's line of argument – demonstrates how we can be considered morally responsible for our acts, it does not adequately resolve the issue of freedom (Fischer 2007, pp.71-77). Our position on freedom could likely lead to what Fischer refers to as a "dialectical stalemate", as it is dependent on which definition of freedom we follow. Either we demand that freedom depends on being able to do otherwise, or we accept that freedom is based on consistency between motives and acts. The tension between these two positions is irresolvable, according to Fischer.

To return to the example of the unchained prisoner, we can say that she chooses to read the book voluntarily, but that this act is not expressive of her free will. This claim is supported by the neural evidence. When we perform a task we perceive as being voluntary, neural regions associated with voluntariness will be activated. If a prisoner walks to the other side of her cell to pick up and read a book, we would expect to find that these neural regions would be activated, and she could report her act as being voluntary. At both the phenomenal and empirical level, then, we can judge the act as being voluntary. This is not to say that she stays in her cell reading of her free will, but that this act is nevertheless voluntary. Olsaretti, in arguing against the rights-based definition of voluntariness, claims that we cannot consider a prisoner as staying in a cell voluntarily (Olsaretti 1998, p.59). On Olsaretti's view, voluntariness is more demanding. However, at the neural level, to claim that an act is voluntary is not to say that it is expressive of an agent's will, as understood through her overarching desires and motivations in life. An act is voluntary if the agent feels it was voluntary and neural networks associated with voluntary action are active in its performance. Considering such acts as being voluntary, though, does not commit us to also claiming they are acts of free will, as they are not necessarily expressive of the agent's will.

In the case of Jones, however, his act does reveal his will. Accepting that Jones acts of his free will, though, forces us to commit to the compatibilist definition of freedom, as Fischer argues. If instead we argue that Jones committed the act voluntarily, but not necessarily of his free will, we can still hold Jones responsible, without committing to either definition of freedom. While we might hold Jones responsible if his behaviour can be seen to be voluntary, the deeper problem of his character requires separate consideration. As Roth states, we have no choice over the neural activity influencing our acts in everyday life. Even when we deliberate, the neural processes influencing deliberation occur largely unconsciously (Roth 2010, p.239). Though we can act voluntarily, we cannot freely and consciously choose the desires and motivations that inform our characters. We can accept Frankfurt's arguments for moral responsibility while still considering this a problem. Though we may be able to hold a person responsible for her acts if we accept this view, we cannot say that she has much ability to control her own thought processes.

To sum up, the neural evidence should force us to reconsider some of the claims of compatibilism. Hume's conception of freedom tells us about external conditions, but little about the constitution of the internal will. Any agent who is free of chains can be considered to be acting of her free will. This conception of free will does not account for the various inner motivations we may hold. Thus, our everyday voluntary actions would be considered free, whether or not they are in accord with our internal motivations. Frankfurt's account of free will does, however, account for these inner motivations. For Frankfurt, it is our second-order desires that enable us to act according to our free will. As I argued, however, though we may hold people responsible for their voluntary actions on Frankfurt's view – providing their first-order desires are in accord with their second-order desires – a problem is left unresolved. We cannot formulate our own second-order desires. The neural processes involved in deliberation over actions occur outside of conscious control. There is no way in which an agent can consciously formulate her own character or her own second-order desires, or control the way in which these aspects of her character influence the actions she chooses to make. Whether or not we accept Frankfurt's formulation of moral responsibility, we can still accept that this inability poses a problem. Whatever our stance on free will and moral responsibility, there is little sense in which a person can be said to be in control of her own thoughts.

7.8 Conclusion

Libet thought his discovery – that the readiness potential preceded conscious awareness of the intention to act – challenged traditional notions of free will. Libet did not argue against the truth of free will, however. Instead, Libet thought that the veto function allowed for our free will. Because we can veto the impulses toward action encoded in the readiness potential, we have conscious control over our bodily movements. This allows for a conception of free will that is libertarian and dualistic: independent of deterministic laws and founded on a distinction between the mind and body.

In philosophy, both Libet's conception of free will and the assumptions made based on his experiments have been challenged. Throughout this chapter, I have assessed the merits of these challenges. Some, such as Mele's arguments for the importance of motivations, fail to establish a more cogent interpretation of the data, as the basic premise of Libet's argument is retained. Mele attempts to show how our conscious motivations can have causal functionality. For this argument to work, we would need to accept that consciousness is somehow independent of other causal laws. None of the later empirical evidence within neuroscience supports this view. In the work of Roth, Haggard, and others, the brain is viewed as working deterministically. Neural regions communicate with one another via the transmission of action potentials, with no single region representing our experience of consciousness, or operating as a command centre that is independent of causation, issuing orders to the rest of the body. Libertarianism becomes incoherent when assessed against understandings of neural activity.

Dennett provides a more convincing interpretation of the neuroscience. Upon Dennett's view, there is no Cartesian theatre, a seat of consciousness from where action is controlled. Consciousness exists instead in processes across time. Roth's work supports this view; the cerebral cortex is not the command centre it was once portrayed to be. Other neural regions interact with each other to create what is felt as a conscious experience by individuals, with each region playing a significant role. While Dennett's interpretation of the data provides a plausible conception of consciousness, it does not, however, tell us much about freedom.

There are several possible compatibilist responses here. None of them, however, are entirely satisfactory. Dennett's own view – that we learn from experience, enabling alternate possibilities in the future – does not tell us at what point we are exercising our free will. If free will evolves in the way Dennett supposes, agents temporally prior to other agents will always have less free will, though we could never be sure when an agent begins to exercise her free will. On Hume's argument, we learn little about the internal states of agents or the realisation of desires. On a Frankfurtian account, though we learn more about the internal motivational states of agents, a problem remains. We do not consciously formulate our own characters and desires. We could accept this compatibilist account of free will and moral responsibility, while still seeing why this is problematic.

There are, then, problems left unresolved in relation to how our will is formed. We do not form our own characters, or our responses to external phenomena in full conscious control of the formation. The level of control we have over our own thought processes is small. If one accepts the compatibilist position, she may maintain a commitment to the truth of free will and moral responsibility. The problem that we hold little conscious control over our thought processes remains, however, unresolved.

Chapter 8 – Rational and Full Autonomy: Considering the Implications of Neuroscience

8.1 Introduction

In this chapter, I assess the implications of the claim made in Chapter 7: we hold little conscious control over our thought processes. In light of these implications, I argue the way in which Rawls posits individuals as becoming autonomous should be revised. Our thoughts are formed by neural processes over which we have little control. Even if we endorse compatibilism – which is not undermined by the empirical evidence in the way in which libertarianism is undermined – we should consider the lack of conscious control over thoughts a problem. In a society containing many unreasonable ideas, we could not expect people to remain uninfluenced by these ideas. Such ideas would hinder the prospects for these people to realise their autonomy in the way Rawls formulates it.

If we lack conscious control over our thought processes, the way in which individuals are thought to realise their autonomy should be revised. There are two forms of Rawlsian autonomy: rational autonomy and full autonomy. Rational autonomy relates to our ability to realise a conception of the good, ensuring our life plans deriving from this good fit with those of others in a socially cooperative world. Full autonomy relates to our political lives; to be a fully autonomous citizen, an individual must refer back to the principles of justice when engaged in political decision-making. As noted in Chapter 3, without fully autonomous citizens, there is no one to realise the values underpinning Rawls' theory of justice. If we cannot rely on citizens to develop their autonomy when left to their own devices, we must seek others means through which to enable the development of full autonomy. This requires us to take a different approach to the way in which citizens develop both their rational and full autonomy from that proposed by Rawls.

Political liberalism is dependent on the existence of fully autonomous citizens who recognise the importance of the values embedded within liberalism. Whether or not such citizens come into being is dependent on the ways in which their moral psychologies are developed. Considering the implications of neuroscience should lead us to think about these processes of development. Through such consideration, it should be recognised that the conditions that enable the development of full autonomy should be prioritised, rather than prioritising certain legal rights and liberties. The defence of such rights and liberties is dependent on citizens who are motivated to defend them.

I return to the claim that we hold little control over our thoughts in 8.2, spelling out the implications. We have, I argue, little control over our conception of the good if the claim is true. Rational autonomy is assessed in 8.3. I explore how the considerations following from the implications of neuroscience may alter the process of reflective equilibrium in the original position, forcing us to slightly revise how we consider rational autonomy. Following this, full autonomy is assessed against the data in neuroscience in 8.4. Without fully autonomous citizens motivated to defend basic rights and liberties, political liberalism is left unstable. Rather than leaving citizens to determine their own good, I argue we should ensure the development of full autonomy across society. This has implications for the philosophical and religious doctrines existing within a society; many such doctrines may disagree with the implications drawn from the empirical data. In 8.5, I note that the implications of neuroscience may conflict with certain philosophical and religious doctrines. However, as I argued in Chapter 4, if people object to political conclusions reached in light of science, it must be their responsibility to offer a better explanation of the science.

8.2 The Lack of Conscious Control

If we hold little conscious control over our thoughts then, I argue, the way in which we form a conception of the good will largely be beyond our control. In this section, I argue why this is the case: if we hold little conscious control over our thoughts, we also hold little control over our character development or the conception of the good we come to hold.

According to Libet, we cannot be held responsible for our thoughts (1999). It may be thought that, according to common-sense, we do not hold people responsible for their thoughts anyway. Though we might be shocked to hear that a friend had thought about murder, providing this person did not act on her thoughts, we may not think much of it. Unwelcome thoughts often slip unbidden into our minds. We do, however, hold people responsible for what derives from their thoughts. If a person seems only to think of herself, we may consider this person selfish or egoistic, and seek to avoid her. According to the implications of neuroscience, however, the selfish person had little control over the formation of her character. We may hold her responsible if her selfish acts overstep the bounds of what is considered reasonable behaviour – that is, we may alter our behaviour in response – but we cannot consider her as possessing the ability to alter the thought processes involved in the formation of her character. There will be, then, implications that follow from a person's acts and behaviour, as such acts and behaviour provide us with reasons for responding in certain ways. However, we should not imagine that a person holds the capacity to alter her thoughts and character. For the Rawlsian, we could not hold that the unreasonable person consciously determined her unreasonable character.

As I argued in Chapter 7, while Frankfurtian moral responsibility holds at the level of action, it is not applicable at the level of thought. At the level of thought, whether such consistency holds is essentially a matter of luck; that is, it is outside of human control. Whether one thought remains unchanged from moment to moment is due to the activity of different regions of the brain, with no 'command centre' from where this is controlled. We can imagine a scenario in which a thought enters our conscious awareness due to the habitual workings of the basal ganglia, yet is altered by the triggering of a memory within the amygdala (see 6.2 for an explanation of these nuclei). That this alteration took place is not due to agential control, but a matter of luck within neural activity, over which we have little control. Frankfurtian moral responsibility requires more than this. Within action, there are two clear aspects: the character and the act. Within the relationship between the character and the act, there is a centre of agential control. Central agential control can be attributed if there is consistency between both the character and the act, and the act can be said to be voluntary. No such central agential control exists within thought processes. The relationship between two neural processes is a matter of luck, given that there is nothing to control this relationship.

Character primarily belongs, I argue, to the domain of thought. If we hold little capacity to control our thought processes, it cannot be imagined that we have much control over the characters we come to hold. Our characters are mentioned here only in relation to what is relevant to moral and political matters; I do not offer a full account of character formation or what it means to have a sense of self. The sorts of thoughts we think and the ideals we hold are constitutive of our moral character. From a neural perspective, we have little conscious control over how these thoughts and ideals influence our character formation. Though we may alter our initial ideals in response to external events, or the formation of new thoughts within the brain, we do not consciously formulate our responses to events or the new thoughts that lead us to reject our original thoughts. As we do not consciously initiate our own thought processes, we must assume that our thoughts develop through processes external to conscious awareness, and so the external environment and our internal genetics will have played a role in their formation. The ideals and values that are dominant in a society will, therefore, influence the thoughts, and thus also the characters, of people living in that society in ways beyond the control of those people.

Though I argue against this contention, it could be claimed that both thought and *action* influence character development, and that, therefore, we should view character development as being an equal product of each.[1] For instance, a person who previously hated tennis may come to love tennis due to taking part in a game. In this way, the act of playing tennis has played a part in this person's character development. While this is true, thought and emotion are crucial in this development. Thought is always a necessary component in character development, whereas action is only supplementary. Without a positive emotional reaction to the act, followed by a conscious awareness of this reaction, there could be no development: the act detached from thought could not lead to character development. Whether we respond positively or negatively to an act is due to neural processes. We cannot dictate such processes, and though someone with a particular strong aversion to tennis may still claim to hate the game, she could not have prevented herself from enjoying playing the game if her enjoyment resulted, for instance, from a sudden release of dopamine. Or perhaps someone with a previous dislike of tennis wished she could change her opinion, and so, through what seems to her as an effort of will, begins to enjoy tennis. Whether or not the effort of will succeeds is subject to the occurrences of future neural activity, something over which she has little control. A sudden recollection of her previous dislike triggered by activity in the limbic system may cause her to abandon her efforts to change her opinion. Our characters, consisting of our likes, dislikes, desires, fears, and many other aspects, are formed through our thoughts, and our reflections on experiences. We have little conscious ability to shape these thoughts and reflections. Thus, while action may play a role in character development, it is supplementary to thought. For this reason, I argue we should consider character as belonging primarily to the domain of thought, rather than action.

[1] This is why many argue for a moral education based on performing certain activities; it is the type of activities a child takes part in that enable her to develop the character she develops (see Nucci, Narvaez, and Krettenauer 2014).

What Rawls describes as our conceptions of the good can, I argue, be thought of as connected to our characters. A person whose deep sense of Christian faith leads her to live a life devoted to charity will likely possess a character of which faith and altruism are important components. We could imagine a person living a similar life, but rather than out of a sincere belief in the importance of faith and charity, she leads such a life to receive public acclaim. Despite the insincerity involved in the second scenario, however, in both cases, the conception of the good follows from the person's character. Both accept conceptions of the good that are in keeping with their characters, whether sincere or not. The conception of the good, and the acts that follow from it, work to uphold the person's sense of character. While the first person's conception of the good is living the life of a good Christian as she sees it, the second person's conception of the good is living a life to receive public praise. In the former case, acts of charity strengthen the person's faith, while in the latter case, such acts work to enhance her chances of receiving acclaim. The notion of a person living with a conception of the good entirely at odds with her character is incoherent. Whatever the nature of the connection between these two concepts, it is an essential component in the development of a conception of the good.

If we have little control over our thoughts, along with our characters, which we possess largely as a result of our thoughts, then neither can we be said to consciously control the way we formulate our conceptions of the good, which are intimately tied to our characters. The person with a deep sense of Christian faith will likely have lived a life in which she was moved by learning certain aspects of Christianity, or by some religious experiences. Her reaction to such experiences is not within her control, as was argued earlier. Alternatively, she may have never questioned the faith with which she was raised. Again, she holds little control over this. The fact that certain thought processes were never triggered is not something we can expect someone to alter. Perhaps the world in which she lived never offered her experiences which challenged her beliefs, or perhaps her beliefs were so strongly embedded that no experiences would have altered her beliefs. Either way, there is little control exercised over the conception of the good a person comes to hold.

To conclude, there is no part of the brain that is detached from other parts, or that allows for decision-making that is independent of other neural events. The path between our thoughts, intentions, and desires that leads towards action is deterministic. The thoughts themselves are not subject to much conscious control as there is no regulative centre over the neural processes that form our thoughts. The connections that persist between our thoughts is largely due to luck – where luck is understood as what is beyond human agential control – as there is no centre of conscious control over these relationships. Thus, the extent to which we are in control of the characters we become is limited. As the conceptions of the good we hold will be, to a large degree, dependent on these characters, we also hold little control over the way in which we choose our conception of the good.

8.3 Rational Autonomy and The Original Position

In light of what was argued in the previous section – that we hold little conscious control to formulate a conception of the good – I argue here that we should revise how an individual becomes rationally autonomous. Rational autonomy plays a specific role within the original position in relation to how we are to settle questions regarding human nature, conceptions of the good, and the plans of life people live in accordance with. If parties consider the implications of neuroscience, the decisions reached in the original position would be altered. Rather than assume the moral powers and self-respect to be an innate component of human psychology, parties would instead realise that the rational autonomy of those they represent is dependent on the conditions in which they live. In particular historical moments – such as ones involving intense opposition between religious movements – the commitments entailed by Rawlsian moral psychology would be absurd. Looking instead to the empirical data offers a better basis on which to comment on human nature. It should not be assumed that the individual possesses two moral powers. Instead, if individuals are to develop the type of moral psychology necessary to support the values of political liberalism, this development needs to be fostered by society.

Rawls admits that in certain historical periods, attaining an overlapping consensus of reasonable doctrines would be an impossible task (Rawls 2005, p.126). This was true during the religious wars between Protestants and Catholics in the sixteenth century, to refer to Rawls' example (2005, pp.148-149). In such a time, it would be absurd to imagine that there could be common agreement on what is just, or that all parties would wish to seek fair terms of cooperation with one another. What Rawls does not acknowledge, however, is that it would also be absurd to imagine that people in such a time would possess the necessary moral psychology that underpins political liberalism. It could not be supposed that Catholics would wish to form a conception of the good that cohered with those of Protestants, or vice versa. Neither would people possess any motivation to develop a sense of justice which was premised on being fair to the opposing party.

From the perspective of the original position, given that the veil of ignorance removes our knowledge of the way in which historical circumstances have affected the social world in which we live, parties could not be sure whether the citizens they represent live in such a time. The realisation of justice as fairness, and with it, the securing of citizens' interests, is dependent on people possessing the appropriate moral psychology. Therefore, parties in the original position have an interest in securing the conditions necessary to allow for the development of this psychology.

In part, Rawls' conception of rational autonomy relates to the parties' ability to achieve this. Rational autonomy is, in essence, the ability to develop the moral powers, living a life in accordance with these powers. Both citizens and their representatives are considered rationally autonomous (Rawls 2005, pp.72-77). Parties are rationally autonomous on two conditions: first, they are able to decide on fair terms of cooperation, with no principles of justice previously determined to guide these decisions, and, second, they recognise that the citizens they represent have higher-order interests in realising their moral powers. There are also two conditions for citizens to be considered rationally autonomous: they are able to pursue a conception of the good within the limits of political justice, and they desire to realise their moral powers.

As the original position is a hypothetical situation, we can suppose that deliberations within are not subject to the limits of physical laws. We should not, however, apply the same standard to citizens. Thus, we can apply the implications of neuroscience when considering how citizens determine their rational autonomy, but not to parties in the original position. Recognising, then, that the citizens they represent will be constrained by the implications of neuroscience, I now turn to examine how this might affect deliberations.

William Simkulet (2013) argues that if hard determinism was considered to be true within the original position, we would come to radically different conclusions from those drawn by Rawls. Here, I am not supposing the truth of hard determinism, but attempting to establish the implications of neuroscience. We will consider these implications here as background theories within reflective equilibrium. In trying to reach reflective equilibrium – a state in which, as explained in Chapter 4, principles cohere with judgements when assessed against one another (Rawls 1971, pp.48-51) – we will assess whether the principles of justice, and the sorts of judgements we make in light of these principles, fit with the implications of neuroscience.

As Richard Arneson notes, Rawls attempts not to establish distributive principles on the basis of desert, but without removing individual agency entirely from his account of justice (Arneson 2008, p.85). Saul Smilansky has similarly noted that Rawls begins a hard determinist, but reaches compatibilist conclusions (Smilansky 2003, p.132). That is, though we are not to distribute primary goods according to desert initially, Smilansky argues that after this is established, Rawls assumes people are morally responsible for what they do with their share of these goods. Thus, citizens hold duties in relation to what they choose to do with their share of the primary goods. As Rawls assumes that citizens possess a certain moral psychology, he imagines that citizens will be motivated to act in accordance with this conception of duty. This is why self-respect is important for Rawls (1971, pp.440-446). We increase our sense of self-respect through realising successful life plans that cohere with the type of cooperative society in which we live. Someone who had no interest in realising such a plan of life, who thus had no desire to realise her moral powers, would not be a "full person", according to Rawls (2005, pp.76-77). Hence there is a motivational force behind self-respect. It is because we possess this type of moral psychology that we desire to act responsibly.

Given parties in the original position do not know what sort of historical circumstances citizens will live in, they could have no certainty that citizens possess this form of moral psychology. They could not suppose that Protestants in the sixteenth century would increase their sense of self-respect through finding that their plans of life cohered with those of Catholics. Parties in the original position would, then, need to be wary of the assumptions they made regarding moral psychology. For this reason, rather than base our understanding of autonomy on assumptions regarding the type of moral psychology people possess, we should instead base any such understanding on empirical data.

From the perspective of the original position, if parties are to secure the interests of those they represent, they have an interest in society being devised so as to inculcate the appropriate dispositions in citizens. There are two reasons for this. First, without these dispositions forming the basis of Rawlsian moral psychology, political liberalism is unrealisable. Second, it is in the interests of citizens to possess such dispositions if they live in the type of society imagined by Rawls. If they fail to develop such a moral psychology, they will be at a significant disadvantage; others would not wish to cooperate with them, and their way of life would face constant obstacles. However, it can neither be assumed that citizens naturally possess this form of moral psychology nor that they will be motivated to act responsibly. If the implications of neuroscience are true, we hold little control over the development of our tastes. Rawls expects us to keep our tastes within the confines of what we can reasonably expect our share of the primary goods to be (Rawls 1982b). The person who develops an unreasonable taste for expensive goods could do little to alter the development of this taste. Rather than leave citizens to determine a conception of the good through their own capacity for reason, it should be recognised that a citizen left to her own devices cannot be guaranteed to develop her thoughts, character, tastes, and values in accord with the values of political liberalism. Her conception of the good could, then, be at odds with the values of political liberalism. Thus, parties would ensure that the basic structure of society was devised so as to motivate a person to develop a reasonable conception of the good, rather than assume citizens will be motivated to do so due to their innate moral psychology. It cannot be assumed that a person will develop her rational autonomy in accordance with the values of political liberalism if left to her own devices.

Rawls recognises that without the publicity condition being met, we could not expect citizens to act as fully autonomous citizens. That is, without knowing what justice is, and what requirements it places on citizens, we could not assume citizens would act to realise justice (Rawls 2005, p.78). However, citizens are expected to realise their rational autonomy under different conditions. As citizens are to determine their own conceptions of the good, the structure of society could not influence the development of a person's rational autonomy in the same way. A particular conception of the good being embedded in the basic structure of society would undermine this task. Thus, no one way of life could be encouraged at the expense of another. Thus, though Rawls assumes that in a just society, citizens would be motivated to act as fully autonomous citizens, the life of a fully autonomous citizen – a politically engaged person who acts in accord with the principles of justice across her life – could not be promoted at the expense of other ways of life. However, if it is recognised that citizens hold little conscious control over their thoughts, it would also be understood that leaving a citizen to determine her own good entails great risk. In a society containing many ideas that contradict the values of political liberalism, we could not assume she would develop a conception of the good that is reasonable.

As argued in Chapter 3, justice as fairness requires fully autonomous citizens. However, if people are to become fully autonomous citizens, their rational autonomy needs to develop in accordance with the demands of full autonomy. Before realising their conception of the good, citizens must have developed a particular form of moral psychology. If people hold little control over their thought processes, then we cannot expect people to necessarily develop their moral psychologies in the appropriate way. When a doctrine is deeply embedded within a person's sense of history, such as many religious doctrines, or when a doctrine appeals to certain aspects of a person's self-interest, the doctrine may possess a strong motivational force. For the realisation of justice as fairness, the motivational strength to endorse the principles of justice needs to outweigh that of other comprehensive doctrines. If parties cannot rely on the innate moral psychology of citizens to fulfil this role, they will need to look for others means. Rather than make assumptions about the state of moral psychology existing within a society, we should instead think about the way in which society affects how citizens develop their autonomy. Thus, the appropriate mode of psychological development must be encouraged through the structure of society. It cannot be assumed that citizens are rationally autonomous and will develop the necessary psychological commitments of their own accord. These commitments must be fostered.

To conclude, considering the implications of neuroscience within the original position leads to revised conclusions. Rather than imagine citizens possess a moral psychology that enables them to become rationally autonomous, and that self-respect ensures that citizens are motivated to do so, parties in the original position will instead recognise that citizens have little control over the ways in which doctrines influence their thoughts. We cannot depend on citizens possessing the form of moral psychology Rawls assumes them to possess. Instead, the way in which society influences the development of a person's autonomy should be reconsidered. These influences should be structured in such a way to enable a person to become autonomous.

8.4 The Realisation of Full Autonomy

In this section, I provide a brief recapitulation of full autonomy, before examining how the implications of neuroscience should lead us to recognise how integral full autonomy is to political liberalism, yet how easily it is undermined. Considering neuroscience within the original position forces us to think about how the development of our autonomy is influenced by society. If citizens hold little conscious control over their thoughts, their autonomy will be largely determined by external forces. This being the case, society must be devised so as to ensure the development of fully autonomous citizens, without whom, political liberalism is unrealisable. We cannot rely on citizens becoming autonomous through exercising their right to freedom of conscience. If society is to be inhabited by fully autonomous citizens who uphold the principles of justice, the development of the necessary moral commitments must be the focus of the basic structure of society. As Rawls recognises, it is only with the publicity condition being met – citizens recognising that the principles of justice are reflected in institutions across society – that a person could be considered fully autonomous. This, I argue, is not enough to ensure a person becomes fully autonomous. To ensure the development of full autonomy, the value of autonomy must be promoted across society; more is required than recognition and understanding.

When a citizen is fully autonomous, she recognises that the principles of justice are just, complies with the principles, and acts *from* them (Rawls 2005, pp.77-81). Rawls stresses that this is not autonomy as an ethical conception, as in Kant or Mill, where the promotion of autonomy determines other aspects of moral theory. Instead, it is only applicable to the political realm. When deciding matters in the political realm, the fully autonomous citizen looks to the principles of justice to determine what should be done. Though full autonomy is grounded in our rational autonomy – our ability to form a conception of the good and realise our moral powers – whereas rational autonomy is applied to the original position, full autonomy is only realised outside of the original position. It is not applicable to parties but to citizens. To satisfy the conditions of full autonomy, the publicity condition must be met (Rawls 2005, p.78). This means that justice as fairness and the obligations that derive from it are well-known. They are embedded in public institutions and citizens are made aware of what justice as fairness demands of them. From this level of awareness, citizens are able to act to realise full autonomy in their political lives.

If the implications of neuroscience are considered, it will be recognised that full autonomy is more vulnerable than Rawls imagines. There are two problems here. First, as noted in 8.3, Rawls assumes citizens possess a certain moral psychology. It is because they possess such a moral psychology that they are motivated to act as fully autonomous citizens. Second, Rawls imagines citizens have intellectual powers that enable them to exercise their right to freedom of conscience. Though Rawls does not state this explicitly, as citizens are motivated to act as fully autonomous citizens (Rawls 2005, p.78), we can rely on citizens to overcome the influence of ideas that contradict justice as fairness. When doctrines are presented within the public culture – political libertarianism, religious fundamentalism, or racist ideas that reject the basis of moral equality, for example – the fully autonomous citizen rejects these doctrines, as with the publicity condition being met, she recognises that it is in her interest to do so. Thus, though Rawls does not put it in these terms, if a citizen is motivated to act in accordance with the principles of justice, she will reject doctrines that do not cohere with the principles.

As I have argued, we should not make these assumptions. Instead, if citizens are to behave in this way – acting to support the principles of justice and uphold the constitution – the structure of society needs to be devised to motivate people to do so. Rawls appears to assume that recognition and understanding of the principles of justice is sufficient to ensure a commitment to them. However, if citizens have little control over their thought processes, a deeper commitment to the principles is needed; that is, an internal motivation to endorse the principles of justice, not just an acceptance of the rules stemming from a understanding of the benefits they produce. Without such a commitment, citizens may be influenced to act against the principles of justice by alternative doctrines. If citizens were in full conscious control of their thoughts, then it could be imagined that citizens would possess the intellectual capacities to reject opposing ideas. Even without a deep commitment to the principles of justice, recognising that they advance our own interests may be enough for us to remain supportive of the principles.

However, in light of the implications of neuroscience, it cannot be assumed that citizens will commit to the principles of justice only through recognition and understanding. People lack the ability to consciously control their thought processes, and we cannot assume that the majority of citizens will not be persuaded by doctrines at odds with political liberalism. Furthermore, as argued in Chapter 3, reason is secondary to emotion in determining our sense of morality. This being the case, if we are to ensure citizens are committed to the principles of justice and act as fully autonomous citizens, it should be ensured that a person's emotional commitments cohere with the Rawlsian sense of justice.

In this way, considering the implications of neuroscience leads to a better means through which to realise full autonomy than does Rawls' conception of publicity. Instead of assuming the possession of a particular conception of moral psychology, through instilling the appropriate moral sentiments in people through the basic structure of society, we increase the likelihood of people becoming fully autonomous. Neural activity being deterministic, a person with a prior deep commitment to the principles of justice is unlikely to be persuaded otherwise.

However, it could be argued that ensuring the development of fully autonomous citizens undermines freedom of conscience.[2] The basic liberties are granted lexical priority throughout Rawls' work. However, there is a tension here between this protection of the basic liberties, and the type of person Rawls wishes to inhabit the well-ordered society he envisages. As Frankfurt thought that a person who had no second-order desires could be described as a "wanton" (Frankfurt 1971, p.11), as such a person would act only on first-order instincts, possessing no secondary desires regarding the desirability of these instincts, Rawls similarly argues that a person without a wish to develop her moral powers would not be a "full person" (Rawls 2005, p.77). In order to realise our full human potential, for Rawls, as for Frankfurt, we must be motivated not only to act on instinct, but to possess a certain moral psychology, which we desire to develop in order to understand and act from principles which we consider to be right and good. People who do not desire to use the freedom granted to them by the rights and liberties in their possession are, for Rawls, not developing their true human nature. Once we possess legal rights and liberties, we must, Rawls posits, think and act in certain ways relative to our moral psychology if we are to become "full persons". The rights and liberties alone are not, however, sufficient to guarantee people will wish to become such persons.

The basic liberties are, then, primarily of importance insofar as they allow for the realisation of full autonomy; outside of this relatively narrow scope, the basic liberties can be a threat rather than a benefit to the realisation of a well-ordered society. Rather than prioritise the basic liberties, which do not necessarily guarantee the development of full autonomy, we should instead prioritise the means that allow for autonomy's realisation: moral education, publicity, and the material resources that enable a person to become fully autonomous. This is not to say that such citizens will not care about basic liberties. Instead, it is to say that if we are to live in a society in which people care about liberal values, we must ensure society is structured so as to enable people to become fully autonomous.

[2] See Andrew Murphy's article 'Rawls and a Shrinking Liberty of Conscience' (1998). Murphy argues that while Rawls considered political liberalism to be an extension of liberty of conscience, it is instead a retreat from this liberty.

This leads us to reassess the importance of freedom. Rawls claimed that, among people living in modern societies, there was a general preference for freedom over other values (Rawls 1971, p.28). People would not be willing to trade their freedom to live in a more egalitarian society, according to Rawls. However, none of the values Rawls considers important would be realisable without individuals motivated to realise them. Under a scheme of legal rights and liberties which protected the negative liberty of the individual, but did not necessarily motivate the individual to accept liberal values, we could not expect to find such motivations in individuals. If people have little conscious control over their thoughts, we should question the priority of Rawls' first principle of justice. The scheme of liberties this principle offers would lead to a society of rationally and fully autonomous individuals only under certain other conditions. It is these conditions we should prioritise.

8.5 Neuroscience and the Plurality of Doctrines

Opening up the scheme of political liberalism to the consideration of science and metaphysics leads us to the risk of conflict between competing doctrines. If we demand that constitutional matters be subject to the limits of scientific knowledge, some may worry that such knowledge conflicts with how they perceive the good, or their account of truth. Perspectives on the question of free will found within religious doctrines may not cohere with the empirical evidence. People of faith may worry that the implications of neuroscience are supportive of other doctrines over their own. However, neuroscience and its implications do not necessarily conflict with differing reasonable doctrines. It is possible that most comprehensive doctrines could accommodate these implications without having to revaluate other aspects of the doctrine. Nevertheless, those whose faith is premised on metaphysical libertarianism may be concerned by the implications of neuroscience, as their faith does not accommodate certain empirical accounts of human nature. Such people may object to the political conclusions premised on the empirical data, or wish to propose alternative conclusions. I argue that if such conflicts arise, it is for those whose doctrines conflict with scientific knowledge to show either that the scientific knowledge is wrong, or that their interpretation of the knowledge is the more accurate account. Following this, I explore a problem I claim is more troubling: ensuring the development of full autonomy reduces a person's motivation to accept other doctrines. If a society was structured so as to ensure the development of fully autonomous citizens, those with religious faith may worry that there is nothing within the basic structure of society to motivate people to endorse alternative doctrines.

While it seems unlikely a modern society could be deeply divided by the perspectives people took on the question of free will – the doctrinal conflicts in modern societies tend not to be premised directly on the metaphysics of free will – there are positions on free will within religious doctrines. Christian theologians such as Saint Augustine recognised a similar problem in Christianity to that of the tension between free will and determinism; how to reconcile free will and moral responsibility with divine omniscience and predestination (Augustine 1950, pp.152-156). Only beings before the fall from grace, for Augustine, possessed the capacity for unhindered free choice, Adam and Satan both freely choosing to sin (Augustine 1887; Rist 1969, p.433). Freedom consists only in the freedom to sin; when we act in accordance with the good, we do so due to divine intervention. It is, then, through God that we are able to act from the good. From the perspective of metaphysical libertarianism, Augustine grants us little room for free will, as our ability to do otherwise is dependent on God's will.

Thomas Aquinas, on the other hand, is an incompatibilist, according to Robert Kane, influenced here by Eleonore Stump (Stump 1996, p.75; Kane 2000). The will, for Aquinas, is the primary mover, according to Stump (Aquinas 1947; Stump 1997, pp.578-579). Though the will receives advice from the intellect, the intellect, while it can influence the will, does not perform the executive role during decision-making. The will, receiving information on the external environment, and advice as to the best course of action given the conditions, is responsible for determining its own constitution. Though the intellect is influential in this process, the will is also in control of the intellect. The will wills toward certain objects and not others; it can also will itself. In this sense, the will, acting as the primary mover of all other aspects of the human body and psyche, is free. Against Aquinas, Martin Luther held that the will was not free (Luther 1525). Luther's view is closer to Augustine's, though while Augustine still regarded the will as necessarily being free, Luther considers predestination and free will incompatible. For Luther, salvation came through Christ, not through the individual's capacity to choose salvation for herself.

Perspectives on free will are similarly divided in other religions. William Watt (1946) writes that it has been commonly believed that Christianity allows for free will while Islam insists on predestination, but that this view underplays the complexity of the issue within each religion. As there are stark differences between the Augustinian, Thomist, and

Lutheran accounts of free will, there are similar differences between different schools of Islam on the matter of free will. Other scholars of Islam, such as Radwan A. Masmoudi (2003), stress the importance of free choice within Islam. Masmoudi posits that, according to the Qur'an, religion cannot be forced upon a person; a person must come to believe through their own free choice.[3]

While this is not an exhaustive account of the perspectives on free will within religious thought, it does demonstrate the diversity of beliefs on the matter. As metaphysicians are divided between various accounts of libertarianism, compatibilism, and hard determinism, theologians are divided between belief in predestination, and belief in individual responsibility and the freedom of the will.

A concern here could be that consideration of the implications of neuroscience within political liberalism could be viewed as favouring aspects of certain doctrines over others; those of a religious faith may object to the political conclusions reached. For instance, a Thomist may claim that the implications of neuroscience lead us to reject Aquinas' formulation of the will and favour instead the Lutheran account, as the latter is premised on determinism, while the former is libertarian. Likewise, a liberal Muslim may view the implications of neuroscience as opposing liberal Islam and being more accommodating of the fatalistic teachings of some Islamic scholars. Interestingly, the danger here is of liberal Islam being incompatible with the scientific knowledge considered within political liberalism. It would appear to be unlikely, however, that those theologians who object to such liberal perspectives on religious matters would view the implications of neuroscience as favourable to their own doctrines in a politically liberal state. I now explain how consideration of the implications of neuroscience should not necessarily lead to further conflict between religious doctrines, and a solution for where conflict does arise.

This problem is most pronounced for a position such as the Thomist position, which is dependent on libertarianism; we are free, according to Aquinas, as the will determines itself, and is not formed by prior causes. If we are to demand that public reason be subject to the limits of scientific knowledge, however, the burden of proof is on those of religious faith to show how their doctrines are scientifically plausible, if they wish constitutional matters to be decided through appeal to the tenets of their doctrines. A Thomist who

[3] Amartya Sen also notes the importance of freedom for Buddhism (Sen 1999, p.234).

wished constitutional matters to be premised on the truth of metaphysical libertarianism would need to prove why it is reasonable to assume determinism is false. In the face of the evidence within neuroscience, this seems an arduous task. Though if we are to accept that constitutional matters are to be subject to our most accurate accounts of scientific knowledge, this is the only way in which the issue can be resolved. This is not to say that we must reject such doctrines, however. We need not use the power of the state to censor the views of metaphysical libertarians, for instance, or to prevent people from promoting libertarianism.

Those religious doctrines that reject the truth of free will, such as Lutheranism, certain readings of Augustine, or those in Islam who reject free will in favour of predestination, may be led to one of two positions on political matters. They may, as those Watt describes who shunned medicine (Watt 1946), wish to avoid all human intervention, in which case whether or not someone accepts the faith is a matter for God, not humans. On this view, there is, then, no need for political action to enforce faith as it is not for humans to do so. There is no reason why such a position cannot be accommodated within political liberalism, providing that a person who follows such a faith is willing to support and obey the principles of justice. This sort of doctrine does not attempt to use state power to prevent the acceptance of alternative doctrines, and could therefore be described as reasonable. On the other hand, a believer in predestination may believe in the necessity of the state enforcing a religious doctrine, as do those whom Masmoudi argues against. As acceptance of faith cannot be reached through a person's capacity for choice, faith must be forced upon a person. This, however, is an unreasonable doctrine. Such a doctrine is premised on the use of state power to enforce a comprehensive doctrine across society; reasonable people are thus not able to follow their own reasonable doctrines, as the acceptance of this unreasonable doctrine is forced upon them. The position on free will within such a doctrine is thus irrelevant, and its promotion should be discouraged within a politically liberal state. Therefore, the positions taken on free will within religious doctrines do not necessarily conflict with the implications of neuroscience, and where they do, it must be the responsibility of those whose doctrines conflict with our understanding of scientific knowledge to show that either such knowledge is mistaken, or that theirs is the better interpretation.

A worry of those with religious faith, however, may be that promoting autonomy may reduce motivation to accept faith. As the basic structure of society is to be formulated so as to promote full autonomy, there is, then, nothing within the structure of society to motivate a person to endorse another doctrine. An education system, devised so as to ensure future citizens leave it with an understanding of what autonomy is and how to realise it, would not allow for a child to be motivated towards accepting any one doctrine, other than that of justice as fairness. Rawls thought that this problem was regrettable, but that the importance of endorsing justice as fairness must outweigh that of support for other doctrines (Rawls 2005, p.200). As Andrew Murphy notes, while Rawls claimed political liberalism would allow for the completion and extension of liberty of conscience (Rawls 2005, p.154), political liberalism actually represents a retreat from liberty of conscience (Murphy 1998). The use of public reason means that the religious fundamentalist must either reject or revise her own beliefs to engage in political deliberation. If we demand a commitment to the development of full autonomy, there is a further retreat from liberty of conscience. The influence of alternative doctrines on a person's psychological development must be restricted is she is to become fully autonomous.

Accepting the implications of neuroscience exacerbates the problem. If we are to accept that a person's thoughts are shaped by her environment, that she lacks control over them, and that we must therefore ensure her environment inculcates a motivation in her to endorse justice as fairness, we must also recognise that she is unlikely to be motivated to endorse other doctrines unless she comes to be influenced by them. While she may be influenced by her family's belief in such a doctrine, if all of the institutions of society are designed to promote full autonomy, and to ensure the publicity condition is met, her only encounters with such doctrines will be through her family and within the private sphere. Over time, it seems unlikely that people will continue to endorse doctrines that are not encouraged within the public sphere. There will be fewer opportunities for people to come encounter alternative doctrines if society is devised to promote full autonomy, and fully autonomous citizens will have less reason to seek out alternative doctrines. If a person realises that over time, her children may become less likely to accept her own faith, as they accept justice as fairness as the primary moral doctrine, she may be less inclined to endorse justice as fairness herself. This problem seems to be intractable; while Rawls saw it as a likely but regrettable outcome (Rawls 2005, p.200), if the implications of neuroscience are accepted, it appears to be inevitable and more pronounced. This will be further explored in Chapter 9.

To conclude, where conflicts arise between science and faith in matters relating to constitutional essentials, and those of faith wish to appeal to their doctrines in support of their political proposals, it must be the responsibility of those with faith to show how their doctrines can either accommodate scientific knowledge or provide a better interpretation of this knowledge. What may be of more concern to the person of religious faith – if they accept the implications of neuroscience – is society being structured so as to allow for the realisation of full autonomy. This could undermine the motivation of a person to endorse alternative doctrines. As has been discussed, Rawls recognised this problem and saw it as regrettable. With the structure of society devised towards the end of autonomy rather than freedom – where freedom is concerned only with the protection of rights and liberties – this problem is heightened. Accepting that we hold little conscious control over our thoughts, and, therefore, the necessity of ensuring the development of full autonomy, our hopes of preserving a plurality of reasonable doctrines within modern society may be dashed. In the following chapters, I argue why this should not necessarily be viewed as an undesirable approach to liberalism.

8.6 Conclusion

Rational and full autonomy are not necessarily undermined by the implications of neuroscience. However, the consideration of neuroscience within political liberalism does lead us to revise how we think about autonomy. According to the data within neuroscience, we lack the ability to consciously control our own thought processes. It cannot be assumed that a person left to her own devices will necessarily develop to be fully autonomous, or reject unreasonable ideas, without a prior commitment to justice as fairness. Rather than placing the onus on the person to regulate her own thoughts in accord with the principles of justice, it will be recognised that there is a responsibility on us all to ensure a society contains reasonable ideas, and that our characters develop under the influence of what is just.

Allowing parties in the original position to consider the implications of neuroscience should lead them to revise how they conceive of the autonomy of those they represent. This autonomy, and the self-respect on which it is based, are dependent on certain historical conditions. As parties cannot be sure that these historical conditions exist for citizens, they replace their understanding of self-respect with an assessment of the neuroscientific data. According to the implications of this data, it is recognised that autonomy will only develop if society is devised so as to allow for its development. Moreover, without rationally and fully autonomous citizens, the implementation of political liberalism is impossible. If the stability of a well-ordered society is to be maintained, there must be citizens who are motivated to uphold the values undergirding such a society. Devising society so as to prioritise the development of autonomy would better ensure that such citizens continue to act autonomously in the Rawlsian sense.

Chapter 9 – What this Means for Political Liberalism

9.1 Introduction

In Chapter 7, I argued that we have little control over our thoughts. Our thought processes and the values we hold are largely the result of processes beyond our control. Then, in Chapter 8, I argued that this should lead us to revise how we think about autonomy within political liberalism. We cannot assume that citizens left to their own capacity for free choice will necessarily become fully autonomous. Their conceptions of the good likely being influenced by the environment in which they live, and there being no undetermined space from where they can assess and revise such a conception, means their capacity for rational autonomy – the ability to formulate a conception of the good and live a life in accordance – may undermine the development of their full autonomy – the ability to act in accordance with the principles of justice. Their determining of their good may conflict with their ability to realise their sense of justice. The person who does not realise the appropriate sense of justice cannot be considered fully autonomous. Rather than assume that citizens will act as rationally and fully autonomous agents, we should instead devise the basic structure of society to ensure the development of full autonomy. Without autonomous citizens – people who are motivated to uphold and act in accordance with the principles of justice – political liberalism would not be possible. Thus, from the perspective of the original position, parties, knowing that citizens have little control over their own thought processes, would be compelled to grant priority to the development of autonomy; our thoughts should be guided towards the realisation of autonomy.

In this chapter, I assess what this means for the wider scheme of political liberalism. The basic tenets of Rawls' political liberalism – a well-ordered society regulated by the principles of justice – are not necessarily undermined by the argument. Instead, I argue that the opposition between political liberalism and perfectionism should be reconsidered. For the existence of any well-ordered society, a degree of perfectionism is necessary. As we cannot assume that citizens left to their own devices will necessarily become fully autonomous, and the stability of the well-ordered society is dependent on fully autonomous citizens, society should be devised so as to prioritise the development of autonomy. This means that, in certain cases, perfectionism will be justified. I assess where perfectionism is justified here, arguing in favour of a perfectionist approach to moral education. As I argued in Chapters 7 and 8, we hold little conscious control over our thoughts. Due to this, a perfectionist approach is sometimes necessitated if we are to develop the just society Rawls imagines.

To begin, the problem of unreasonableness is addressed. In 9.2, I argue that proponents of political liberalism, such as Jonathan Quong, underestimate the problem of unreasonable citizens. Quong argues that political liberalism defends the individual against a paternalistic state, which would be the necessary consequence of liberal perfectionism. Against this, in 9.3, I argue that concerns about paternalism are exaggerated. Without any aspects of paternalism within society, we could not guarantee the existence of autonomous citizens, without whom any kind of liberalism is impossible. In 9.4, I assess the problems of this within education. I argue against traditional interpretations of how political liberalism should approach education, on which all that is required is the publicity condition; that children should recognise what the requirements of their role as a citizen. Instead, a deeper moral commitment to fairness is needed. There is a need, then, to inspire a motivation in citizens to act in accord with the principles of justice. To this end, rather than act paternalistically against an adult population, I argue in 9.5 that the moral education of children should be premised on perfectionism. Thus, when a child reaches adulthood, she should be motivated to develop into a fully autonomous citizen.

9.2 Unreasonableness

Defending political liberalism, Jonathan Quong argues that the unreasonable citizen does not pose a threat to the stability of political liberalism. Arguing that Quong is mistaken, I claim that unreasonableness is pervasive in modern societies. Rather than being confined to a small minority of citizens, the problem of unreasonableness applies to many citizens. If, as I argued in Chapters 7 and 8, people hold little conscious control over their thought processes, which are largely determined by the environments in which they live, unless we focus on establishing the appropriate moral sentiments in people, they may come to hold unreasonable beliefs. The citizen who is fully autonomous – who acts from the principles of justice, being motivated to do so – however, is not likely to be influenced by such unreasonable ideas. Thus, there is further reason to prioritise the development of autonomy over other values.

To begin, I offer a brief recapitulation of what Rawls means by reasonableness. There are two basic aspects of reasonableness: proposing fair terms of cooperation and a willingness to recognise the burdens of judgement, the problem that people will often disagree on matters of the truth and the good (Rawls 2005, p.54). A reasonable citizen is, then, a person who other reasonable people would agree to cooperate with as she does not propose terms of cooperation that would harm their interests, nor does she demand others consider her judgements as correct. One important part of being reasonable is that the reasonable person would not wish to use state power to impose their doctrine on the rest of society, suppressing other reasonable doctrines (Rawls 2005, pp.60-61). Thus, the reasonable Hindu respects the reasonable doctrines of others, and while she disagrees with people who are not Hindus, she does not desire a Hindu state. A Nazi is therefore unreasonable as the Nazi desires to impose an unreasonable doctrine – a doctrine that is unfair and one parties in the original position would not agree on as a basis for cooperation – on the rest of society, suppressing reasonable doctrines.

In what follows, I argue that this latter part of reasonableness – the lack of a want to impose an unreasonable doctrine on the rest of society – has been overstated by advocates of political liberalism, while the former aspect of unreasonableness – a willingness to propose fair terms of cooperation – has been overlooked. Many people live their lives on terms that many others would not agree as a fair basis of cooperation. In their descriptions of unreasonableness, political liberals focus on whether the aim of a person is to capture the power of the state, but fail to look at whether a person lives according to fair terms of cooperation. Not all unreasonable people are Nazis, or adherents of other such extreme beliefs. A person may be generally reasonable but engage in activities or endorse beliefs that erode the sense of fair cooperation existing within a society, and, in doing so, undermine the rational or full autonomy of others (examples of this are provided below).

Adherents of political liberalism often downplay the problem of unreasonableness, arguing that there is no need to justify political liberalism to the unreasonable. Following Burton Dreben, Quong argues that the problem of justifying political liberalism to illiberal people is not a philosophical problem. Quong quotes Dreben: "sometimes I am asked, when I go around speaking for Rawls, What do you say to an Adolf Hitler? The answer is [nothing]. You shoot him" (Dreben 2003, p.329; Quong 2011, p.8). For Quong and Dreben, a person such as Adolf Hitler is beyond the scope of political justification. Justifying a political doctrine to such a person is, then, not a philosophical task, nor one that is necessary in devising political principles. Not all unreasonable people are the moral equivalent of Adolf Hitler, however. Many people say, do, and think things that would be at odds with the values of justice as fairness, making them to some extent unreasonable, but not necessarily so unreasonable to be illiberal and beyond hope of convincing. While I agree with Quong that the Ku Klux Klan member and the Jihadist are not likely to be persuaded by Rawlsian ideals, many other citizens, whose beliefs are not so extreme nor entrenched, are to some extent unreasonable, yet justifying justice as fairness to them is essential to the task of political liberalism. If the majority of citizens – who might be unreasonable to a minor degree – cannot be persuaded to fully endorse the principles of justice, then political liberalism is left highly unstable.

To offer a brief sketch of the problem here, I identify several sorts of people who could be deemed unreasonable but who should not be outside the scope of justification. The cases that follow should be considered from the perspective of parties in the original position who recognise the implications of neural determinism.

David is a successful journalist who writes on economics. The paper he works for endorses the monetarist approach to economics, and Jonathan writes in support of positions based on such an approach. Believing inflation as being more of a problem than unemployment and economic inequality, he writes in defence of policies that reduce the risk of inflation while increasing the likelihood of inequality. Such policies are at odds with the difference principle, as the inequalities arising from the implementation of these policies are not to the advantage of the worst off. Otherwise, David holds liberal views, and his conception of the good can be termed liberal. He believes in the basic moral equality of all people, abhors racism, and supports religious freedom though being an atheist.

Syeda runs a social media website. While she supports the principles of justice, and will vote and campaign for policies that uphold the principles, users of her social media site can use it to promote whatever political views they wish. Some choose to promote racist, sexist, and homophobic ideas. Nothing is censored on the site. Thus, many people are exposed to these ideas.

Abdul works in marketing for a successful fast food company. His beliefs can be characterised as reasonable. He believes in charity, donating his time and money to charitable causes for which he feels strongly. The conception of the good he holds is centred on this belief, guiding his rational plan of life. Like David, he believes strongly in protecting the basic liberties of all people, and though he is of religious faith, does not wish to impose his faith on others through the use of state power. In elections, Abdul will vote for candidates who recognise justice as fairness. His work in marketing leads to his influence over many other people. As a result of this influence, many people eat unhealthy food from this fast food company and, in doing so, develop eating disorders. Some of these people are children. As adults, their ability to become rationally autonomous – to develop a plan of life in accordance with a reasonable conception of the good – is undermined by eating disorders they developed as children.

Let us now consider these three cases from the perspective of the original position. On considering the implications of neural determinism, parties would not consider these three people entirely reasonable. Reading David's journalism, a citizen may be persuaded to reject the principles of justice in favour of alternative economic principles. After spending time on Syeda's social media website, a person may find herself endorsing views at odds with the ideal of moral equality. Under the influence of Abdul's marketing techniques, a person may have developed an eating disorder that impedes her ability to become rationally and full autonomous. If we lack the ability to effectively control our own thought processes, ideas we find in the public domain may convince us to think or act in ways at odds with the principles of justice. We lack the ability to persuade ourselves of why we should continue to endorse these principles. Without a strong prior commitment to justice as fairness, a citizen may be persuaded to endorse economic principles that do not cohere with the principles of justice, reject the basis of moral equality between all people, or fail to realise a satisfactory life altogether due to the influence of ideas in the public domain. Citizens left to their own devices cannot be expected to overcome such influences and maintain a commitment to the principles of justice if the implications of neural determinism are true.

Parties in the original position recognising this would not view people allowing the promotion of such ideas as being reasonable, as they realise two negative implications. First, these activities could not be viewed as premised on fair terms of cooperation; as the citizens' interests may be harmed through the influences of these activities, parties would not agree to the terms proposed. This is particularly true in the case of the citizen harmed by the influence of Abdul's marketing. Nevertheless, it is also true in the case of citizens influenced by the work of David and Syeda. A person influenced by David's journalism may be persuaded to vote against her own economic interests, thus rendering her worse off.

Someone influenced by racist ideas found on Syeda's website – quite apart from the threat she now poses to the rest of society – undermines her own moral psychology. In the case that society is regulated by an effective sense of justice, such a person may find herself ostracised and unable to live a satisfactory life. It is, then, against a person's own interests to express such ideas in a just society. Second, the stability of society is threatened in the event of the widespread influence of these influences, the sense of fairness being eroded. If David's journalism became influential, with politicians persuaded to adopt monetarist policies as a result, justice as fairness would be undermined. The influence of fascistic ideas spread on Syeda's website could similarly affect justice as fairness.

There may be two objections to this. First, it might be argued that this assumes that a society should be perfectly just, an ideal which is impossible. Second, this appears to be a matter of luck. In the event of the work of David, Syeda, or Abdul having no influence, we would deem them reasonable. I now respond to each objection in turn.

First, while parties should recognise that no society could be perfectly just, seeking to advance the interests of those they represent, parties would recognise the need to reduce the risk of harms citizens face. If the determinism of neural activity was considered to be false, then we could imagine that citizens would be able to guide their thoughts away from negative influences. In the event of a society containing many ideas at odds with the principles of justice, citizens would be able to reject such ideas and continue to endorse justice as fairness. If the determinism of neural activity is considered to be true, citizens are guaranteed no such ability. Rather than attempting to devise a society which is perfectly just in all cases, parties are instead attempting to minimise risks to citizens. Thus, parties would consider the ways in which society could be devised as to promote the development of the citizens' moral powers – the abilities to formulate a conception of the good and a sense of justice. In this case, the activities of people such as David, Syeda, and Abdul would be discouraged.

Second, if neural determinism is true, the thoughts a person endorses are, to a large degree, a matter of luck. What is meant by luck here is not a matter of chance or subject to randomness – if the determinism of neural activity is true, our thoughts are not the result of chance or random occurrences – but *outside of human control*. Whether or not a person successfully influences another person's thoughts, if the aim of this influence is an unreasonable one – the consequences of its success would undermine the terms of fair cooperation – then that person is acting unreasonably. The aim here is to reduce the risk of unreasonableness while accepting that luck plays a role in determining the beliefs people hold. Whether a person is born into a society that largely conforms to Rawls' conception of justice, or one entirely at odds with such an ideal, is a matter of luck. Furthermore, the way in which a person's thoughts are influenced by the ideas and acts that are commonplace in a society is due to luck. In devising society, parties would seek to minimise the chances of ways of life coming into existence that are entirely at odds with justice as fairness. Luck determines whether an act or an idea influences behaviour across society to such an extent that the stability of this society is brought into question. Nevertheless, parties in the original position would seek to minimise the chances of such acts and ideas having negative consequences. Thus, if the consequence of an act could undermine the sense of fairness within a society, this act would be deemed unreasonable, regardless of its success.

Thus, to a degree, David, Syeda, and Abdul are all unreasonable. Their actions undermine the terms of fair cooperation. From the perspective of the original position, we would not agree to any of their terms. None of them is morally comparable with members of the Ku Klux Klan or Jihadists, however. It is likely that they would find the arguments of justice as fairness convincing, but to realise the kind of society necessary for political liberalism, they would need to make adjustments to their ways of life. David has to be convinced that the economic positions that would support justice as fairness are stronger than the monetarist position he currently adopts, and to write articles in favour of such positions. Syeda would need to think about the types of ideas that are promoted on her website. Abdul would either need to switch careers, or persuade his employers that their produce, and the marketing strategies used to promote it, are not conducive to a socially cooperative society in which people are all capable of realising their rational autonomy.

In what follows, I argue that the prioritising of the development of full autonomy offers an effective response to this problem. Fully autonomous citizens will be less likely to act in the ways in which David, Syeda, and Abdul do. If some citizens do act in these ways, other citizens, who are also fully autonomous, will be less likely to be negatively influenced by their acts.

Nevertheless, if the proponent of political liberalism rejects the claim that David, Syeda, and Abdul are acting unreasonably, more extreme views still pose a problem. Someone may hold racist, sexist, or homophobic views without being a committed Nazi or Ku Klux Klan member. Such a person may even be unaware that her views are racist, sexist, or homophobic, yet these views may influence her vote in elections or the policies she prefers. Many people across a society may hold such views. In the case that such views affect their behaviour in the political realm, the sense of fairness within a society will begin to erode, along with the sense of moral equality between people, as political support for such ideals is lost. This problem is exacerbated if the implications of neural determinism are accepted. It cannot be assumed that people will necessarily reject the influence of such views. People may not be aware of the way in which their own views have been influenced by ideas existing within the public culture.

While Quong claimed justifying political liberalism to Nazis was not a philosophical task, and that it should therefore not trouble the political theorist, the same cannot be said of the sort of person described above. Though such people hold views entirely at odds with justice as fairness, their beliefs are not entrenched in the same way that a Nazi's beliefs are entrenched. We cannot say that justifying a political conception of justice to them is not a philosophical task, thus rejecting the need to do so from political theory, as such people may constitute a large minority, or even a majority, of citizens in modern societies. If the political conception of justice is to attain stability, these people must be convinced by justice as fairness. Furthermore, on endorsing justice as fairness, such people must reject their unreasonable views. It must be realised that racist, sexist, or homophobic ideas are incompatible with a commitment to justice as fairness, and that the expression of such ideas within the public culture can influence other people in ways beyond their control. As the influence of such ideas grows, the stability of a just society is undermined, while – due to the implications of the determinism of neural activity – there is little individual people can do to prevent themselves from being persuaded by ideas existing within their culture. Thus, unreasonable people, to whatever extent they are unreasonable, must be considered within the ambit of justificatory reasons. The consequence of claiming otherwise is to accept that as unreasonable ideas spread within a society, a process beyond the control of individual people, we need do nothing to prevent their influence. Thus, stability would be lost.

To conclude, it is not only people who endorse extreme ideologies such as Nazism or Islamism who are unreasonable. Many people believe ideas, or act in ways, that undermine the idea of society as a socially cooperative union. Justifying political liberalism to unreasonable people is not a matter of attempting to convince Nazis of the virtues of liberalism, but of persuading many people who do not hold extreme beliefs that certain things that they do, say, and think are incompatible with the ideals of political liberalism. To realise political liberalism, such people would not only need to be persuaded of this, but also motivated to change their ways of life.

9.3 Perfectionism and Paternalism

A worry of those sympathetic to political liberalism may be that this is paternalistic: that we are determining the good for others as we judge them incapable of doing so for themselves. Rather than demanding people who are to some degree unreasonable to change their ways of life, we should accept that people have to be responsible for their own ends. This means that people persuaded by the arguments of a journalist to endorse economic principles at odds with justice as fairness must be responsible for this. People who developed eating disorders as children due to fast food must be responsible for overcoming such disorders as adults. On the grounds of the determinism of neural activity being true, I argue that this is mistaken. People have little ability to exercise control over their thoughts. Instead of leaving people to their own resources to overcome negative influences and realise their autonomy, society should be structured as to allow people the opportunities to become autonomous. In this section, I argue that concerns regarding paternalism are sometimes exaggerated. A certain measure of paternalism is necessary in the development of the just society Rawls imagines. Quong's argument against paternalism rests on the individual's innate possession of the second moral power. If this moral power is not innate – which, if the truth of neural determinism is assumed, it could not be – then this argument against paternalism fails.

I have argued that if priority is given to the development of full autonomy, the stability of political liberalism will be better ensured, as fully autonomous citizens are motivated to support the principles of justice, and to act reasonably. To allow for the development of autonomy, however, the state must be justified in encouraging some ways of life and discouraging others. The argument here for the priority of autonomy is grounded on the claim that we have little control over our thoughts. If we cannot control our thought processes, nor the ways in which societal and environmental factors have influenced our thoughts, then if we are to become autonomous citizens, the state must ensure the basic structure of society is devised so as to encourage thoughts conducive to the development of autonomy. This means that the state is justified in discouraging some ways of life if those ways of life undermine the ability of other people to develop autonomy.

Before proceeding, I define perfectionism and paternalism, considering how they relate to one another. Martha Nussbaum identifies Charles Larmore as the philosopher who began debate between the 'perfectionist' and 'political' approaches to liberalism (Larmore 1996, p.122; Nussbaum 2011b, p.5). Following Larmore, Nussbaum views the perfectionist approach as a type of liberalism that is comprehensive, as it incorporates a perspective on the nature of the good life that stretches across the entirety of human life, amounting to a comprehensive doctrine. Whereas the political liberal is concerned only with political factors all reasonable people can agree on, the perfectionist liberal's concern is not confined to the political realm. The perfectionist liberal claims that there is a conception of the good life that it is the role of the state to promote. There is an ideal which makes human life valuable – such as autonomy, for Joseph Raz (1986) – and there should be common agreement on its importance within a liberal society (Nussbaum 2011b, p.11). Politics, then, for perfectionist liberals, becomes a matter of promoting values within the state and society that are considered to be true, not one of attempting to identify specifically political values that all reasonable people can agree on.

Paternalism, on the other hand, is an act in regard to another person arising out of a concern for that person. Such an act limits the liberty of the person who the act is made against, according to Gerald Dworkin. As Dworkin notes, this concern could be that person's "welfare, good, happiness, needs, interests or values" (Dworkin 1972, p.65). These acts must be motivated by a concern for some such aspect of a person for it to be counted as a paternalistic act. Interference with a person's liberty that stems from the want of the person interfering to promote her own self-interest does not qualify as paternalism. Quong disagrees with Dworkin in his explanation of paternalism, as he argues that not all paternalistic acts threaten liberty. Offering someone a reward to persuade that person to perform a particular act can be considered a paternalistic act, on Quong's view (Quong 2011, p.75). Instead, Quong argues that paternalism is an act aiming to improve another person's "welfare, good, happiness, needs, interests, or values" that is "motivated by a *negative judgement*" about the other person's capacities to manage these affairs (Quong 2011, pp.80-81). This could be a capacity for reason, willpower, or emotional management. If I prevent you from acting as you wish because I think you are incapable of determining your own good or realising happiness due to your lack of ability to manage your emotions, then I act paternalistically.

In *Liberalism Without Perfectionism* (2011), Quong argues against both perfectionism and paternalism. My intention here is not to dispute Quong's definition of paternalism, but to show that paternalism is not necessarily unjustified. Quong argues against perfectionism, in part, because he argues perfectionism is necessarily paternalistic (Quong 2011, p.73). On Rawlsian grounds, paternalism is incompatible with the view that all people have the second moral power – the ability to form a conception of the good – as someone acting paternalistically assumes that a person is not capable of determining her own good (Quong 2011, pp.100-107). Acting in this way, we are denying the conception of the person lying at the heart of political liberalism. People are free and equal due to their possession of the moral powers. If we deny this by virtue of acting paternalistically, then we deny the very basis of moral equality on which political liberalism is founded.

Explaining his definition of paternalism, Quong uses the example of a drunk driver (Quong 2011, p.83). If I was drunk and went to reach for my car keys intent on driving myself home, but you took them from me, your action would count as paternalistic. Alternatively, if we had agreed beforehand that if I became drunk and decided to drive myself home, you were to prevent me from doing so, then your action would not count as paternalistic. The difference between the two cases is that, in the former, you reject my capacity to determine the good for myself, whereas in the second, I have already conceded that in the case of my being drunk, I cannot determine this for myself, and wish you to do so. Of key importance is who determines what is my good. If someone else decides for me, that person's act is paternalistic. On the aforementioned grounds that this is incompatible with respect for the second moral power – the ability to form a conception of the good – Quong argues that paternalism is wrong.

However, suppose I was to attempt to pick up my car keys while drunk, intent on driving myself home. You may think that this is the wrong thing for me to do as I may, in the least worst-case scenarios, injure myself or get myself arrested, and in the worst-case scenario, kill myself. Nevertheless, it is for me to determine my good, so you let me drive. However, you may also know that between where we currently are and my home there is a busy road, on which I am likely to injure or kill pedestrians or other road users. In letting me drive home, you are letting me determine that my choice to drive myself home is more important than the right of these people to choose the good for themselves – in this case, the good of not being killed by a drunk driver. Choices are rarely made in vacuums in which the consequences of the choice are born only by the maker of that choice; this is particularly true in the making of political choices.

A problem here is that Quong's example leads us to consider possible arguments for intervention that are justified. My choosing to drive while drunk is not a case of determining my own good but of determining what is right. I am unjustified in choosing to drive while drunk as such an act violates what is right rather than what is good. Though it is left to me to determine the good, I cannot determine for myself the right, or what is just. Other examples of paternalism offered by Quong avoid the blurring of this distinction. If I lie to you about the death of your pet as I think you are incapable of emotionally dealing with death, or I think you lack willpower so omit to tell you about a film showing at the cinema so that you can continue working on a novel, then I act paternalistically (Quong 2011, pp.81-82). I determine what is good for you, as I judge you incapable of doing so for yourself. Neither act here blurs the distinction between the right and the good, and the consequences of either choice would be born only by those involved. If we think that paternalism is unjustified due to its undermining of the second moral power, we will see these acts as being unjustified.

Imagining that individuals' possession of the two moral powers is innate, Quong can then argue against paternalism and state what is a just use of state power on this conception of human nature. This, however, assumes that certain aspects of the individual exist prior to the state. If this is reversed – we assume instead that the state is prior to the individual – then this cannot hold. As Feinberg and Narveson posit (1970), in a society in which rights were not guaranteed, people would not develop the sense of morality and self-respect appropriate for life in a liberal democracy. Thus, if this direction of causation is assumed, the possession of the second moral power is not an innate characteristic of the individual, but a result of a certain political arrangement.

If the truth of neural determinism is accepted, then it will be recognised that this latter scenario is the more appropriate assumption. As I have argued, we should not assume that people have such moral powers. Only in certain conditions will a person have developed her second moral power. As we have little conscious ability to consciously control our thoughts, we will only develop our moral powers if we exist in circumstances that enable us to do so. If we are to live in such circumstances, paternalism will, at times, be both justified and necessary. We cannot rely on there being an innate sense of morality within people. If neural determinism is true, a person's sense of morality will largely be the result of external influences beyond her control. An argument against paternalism depending on people holding an innate sense of morality cannot, then, hold. This is not to say, however, that paternalism is justified against a person who does hold such a sense of morality. If a person has developed to act as a rationally and fully autonomous agent, paternalism against such a person would neither be necessary nor justified. Instead, it is to say that paternalism cannot be thought of as inherently unjust based on a conception of human nature. If we are to arrive at a situation in which this conception of human nature holds, then society must be devised as to enable its development. As part of this process, paternalism will sometimes be justified. Paternalistic acts against David, Syeda, or Abdul would, therefore, not be unjustified.

Of course, the problem here is one of implementation. Even if you agreed that paternalistic acts are sometimes necessary, you may find the illiberal idea of the state censoring journalists, social media sites, or interfering with the nature of a person's work disturbing. I focus on the problem of implementation later in this chapter. I turn now back to the relationship between paternalism and perfectionism. As Quong's argument against perfectionism is based on its paternalistic implications, it will not hold if the argument against paternalism fails.

Under Quong's definition of paternalism, the state is paternalistic if it decides it understands an individual's good better than the individual herself. For Quong, the state is not justified in this, as it undermines the second moral power. This is a necessary implication of perfectionism. However, in accordance with the values of political liberalism, I claim that in certain cases the state should be able to claim it has a better understanding of the individual's good. An individual who succeeds in becoming rationally and fully autonomous is living a better life than one who fails. Someone with no interest in realising a reasonable conception of the good – and thus who failed to become rationally and fully autonomous – would not be a "full person", according to Rawls (2005, p.79). In explaining goodness as rationality, Rawls writes that:

> Even if political liberalism can be seen as neutral in procedure and in aim, it is important to emphasize that it may still affirm the superiority of certain forms of moral character and encourage certain moral virtues (Rawls 1988, p.263).

Rawls is perhaps more relaxed than Quong about certain forms of perfectionism within political liberalism. In *A Theory of Justice*, Rawls writes that particular traits must be encouraged, particularly in regard to the sense of justice the citizen holds (Rawls 1971, p.327). He goes on to say that his contractarian doctrine shares similarities with perfectionism, in that each does not primarily focus on the distribution of welfare, and, in this sense, contractarianism is an intermediate position between perfectionism and utilitarianism. Earlier in *A Theory of Justice*, Rawls also appears to be sanguine regarding paternalism in certain circumstances (Rawls 1971, pp.248-250). Rawls claims that parties in the original position would argue that paternalism is necessary to protect citizens from the "weaknesses and infirmities of their reason and will" (Rawls 1971, p. 249). According to Rawls, where a person is threatened by her own irrationality, paternalism is justified, providing paternalism is guided by the principles of justice. Education must also "honor these constraints" (Rawls 1971, p.250).

There is, then, some overlap between the doctrines of liberal perfectionism and justice as fairness. As Rawls recognises, particularly in work prior to *Political Liberalism*, it is necessary to develop certain traits in individuals if justice as fairness is to be realised. Though less is said regarding the development of traits in *Political Liberalism*, I argue that this still holds, particularly if we assume the truth of neural determinism. If we cannot guarantee the individual's possession of the two moral powers, society must be organised in such a way that the development of the appropriate moral sentiments is encouraged. In cases where individuals act in ways that are at odds with such sentiments, we are justified in claiming these people do not have an adequate understanding of their own good.

Thus, the state would be justified in claiming it understands the individual's good better than the individual in certain circumstances. Where a person's good is entirely at odds with rationality and reasonableness in the Rawlsian sense, the advocate of political liberalism is justified in claiming that the state understands the person's good better than she does herself. An individual whose actions impede her ability to develop her two moral powers is not capable of adequately knowing her own good; the state is therefore justified, to some extent, to overrule her in her decision-making. Quong is, then, not justified in claiming that political liberalism requires the state not to intervene in matters pertaining to the good. The thin theory of the good offered in goodness of rationality offers us a standard against which political liberalism can judge an individual's good.

To conclude, if we are to realise the sort of society imagined by Rawls in *Political Liberalism* – a well-ordered society based on terms of fair cooperation, regulated by the principles of justice – we must prioritise the development of autonomy. If any of the values of political liberalism are to be realised, there must be a society inhabited by people motivated to realise them. Those who advocate political liberalism may be concerned that this is paternalistic. As I have argued, our concern for the development of fully autonomous citizens must be greater than worries over paternalism. This does not mean that we should abandon all concerns regarding paternalism. A state that interferes with journalistic freedom will always be a cause of concern to liberals. However, the charge that a particular course of action is paternalistic is not sufficient to discredit the act. Furthermore, it is wrong to imagine the politically liberal state would remain entirely neutral in matters regarding an individual's good. In certain cases, the state would be justified in claiming it understood the individual's good better than did the individual herself.

9.4 The Problem of Education

In this section, I argue that if autonomy is prioritised within the basic structure of society, then this should be encouraged from childhood, rather than attempting to censor or control the acts of citizens. There are three problems I identify here for education within political liberalism. First, people need to be motivated to endorse justice as fairness. Second, people have little conscious control over their thoughts, which are largely a result of external influences. Third, there is the problem of conflicting moral motivations. I explain Rawls' stance on education, before considering these problems for political liberalism. Because, as I argue, we cannot necessarily be assumed to develop the necessary moral sentiments to endorse justice as fairness, something similar to Cohen's egalitarian ethos needs to be the focus of moral education. However, this focus on a certain ethos comes at the expense of the diversity of beliefs within a society. Nevertheless, I argue that if we are to allow for the development of full autonomy, we must ensure that people are committed to the value of fairness. Fairness should be the central focus of moral education. The problems with education are identified here, before the scheme of education I argue in favour of is explicated in 9.5.

There is little said about education or moral development in *Political Liberalism*. Though we are assumed to have two moral powers, little is said about their origin. Though our capacity for both theoretical and practical reason is claimed to be self-originating and self-authenticating, Rawls, following Kant, argues that the moral powers are to be modelled by the constructivist procedure (Rawls 2005 pp.100-104). Thus, we have an innate capacity for reason, and from this capacity, we can construct political principles which would be acceptable to other reasonable people. Through this process, we can understand that people can be regarded as having two moral powers. While Rawls is not explicit – he is attempting not to endorse Kantian transcendental idealism here, as this would be to commit to comprehensive liberalism – this suggests that our sense of morality, and our ability to regard others as having a similar sense of morality, is rooted in an innate capacity for reason.

As noted in Chapter 3, in *A Theory of Justice*, Rawls offers a sketch of the process of moral development that is more fleshed out. Our sense that we are autonomous moral beings is rooted not in authority but in our capacity for reason (Rawls 1971, p.514). If we realised that our sense of morality was rooted only in what had been commanded of us, we would reject or revise this sense of morality. Rawls here appears to slightly contradict his earlier position that we should be neutral between the empiricist and the rationalist approach to moral education (Rawls 1971, pp.458-461). Whereas the empiricist argues that we need to be motivated to do what is right, and that, therefore, the role of education is to supply these motives, the rationalist argues that these motives are not necessarily missing, and that education should allow for the free development of a person's innate emotional and intellectual capacities. While Rawls claims that each approach has its merits (Rawls 1971, p.461), and that he need not choose between the two, in what follows, Rawls aligns himself more with the rationalist approach. Thus, he argues that it is through the love between parents and their child that the child learns to understand the morality of authority (Rawls 1971, pp.462-467), and it is through entering into the social world that the child begins to learn about the importance of cooperation and the morality of association (Rawls 1971, pp.467-472). Finally, while for a time the child's sense of morality is premised only on an understanding of what others approve and disapprove, as the child matures, she comes to understand and attach herself to moral principles for their own sake (Rawls 1971, pp.472-

479). A person does not act in accordance with moral principles merely because her social group approves of her doing so, but because of her own commitment to these principles. It is here that Rawls' view of moral education echoes those of Piaget and Kohlberg, which, as Rawls himself notes (1971, p.459-461), belong to the rationalist tradition. There is an innate sense of morality within the person that ought to be developed through education. Eventually, the person will develop to understand and act from moral principles due to an internal motivation to do so. There is, then, no need to supply these missing motives, as the empiricist would claim.

As Rawls omits an explanation of the process of moral development from *Political Liberalism*, it is unclear whether the doctrine of political liberalism accepts the process sketched in *A Theory of Justice*. However, as Rawls continues to state that we can locate the origin of reason within the individual's moral consciousness (Rawls 2005, p.100), it would appear that the rationalism of *A Theory of Justice* is not entirely abandoned. In what follows, I assess the problems of this for a process of moral development within political liberalism, if the end is to be a well-ordered society inhabited by rationally and fully autonomous citizens.

As I have argued, a consideration of neuroscience within political liberalism should force us to reconsider the place of autonomy in political liberalism. We exercise little conscious control over our thoughts; what we come to think and value is largely determined by processes outside of our control. Thus, we will only develop autonomy in particular contexts. Autonomy is of vital importance for political liberalism, however. If we are to realise a well-ordered society based on terms of fair cooperation, there must be autonomous citizens motivated to realise this society. Further empirical problems are revealed in Kohlberg's research, as noted in Chapter 3, along with some of the research done on moral development following Kohlberg (Gibbs 2013). The fifth and sixth stages of moral development – in which the individual considers the importance of the social contract, and what has been decided to be right by the whole of society, following which human behaviour can be guided by an understanding of abstract universal principles (Kohlberg & Hersh 1977) – are rarely reached by individuals. This was further demonstrated in subsequent studies (Harkness et al 1981; Colby et al 1983; Mason & Gibbs 1993; Gibbs et al 2007). As Bill Puka puts it, there is no "empirical basis" for stage six (Puka 1990, p.182). Even those sympathetic to Kohlberg's approach concede that stage six is a theoretical aim rather than an empirical fact (Habermas 1990; Gibbs et al 2007).

Instead, most individuals remain at the conventional level of morality, in which individual behaviour conforms to the norms of the given social order the individual lives within (Gibbs 2013). Furthermore, Haidt's research appears to show that it is emotion rather than reason that guides the individual's understanding of morality (Haidt 2007). If reason plays any role, it is largely to justify the initial emotional reaction. Thus, we should not consider individuals as attaining rational and full autonomy with ease. The development of the two moral powers, following which the individual will be guided by an abstract understanding of the right and the good, is not something we witness in those without an understanding of moral philosophy.

This is not to say that Rawls necessarily requires citizens to understand and act from abstract moral principles. Providing a citizen does not act against the principles of justice, and is motivated to support the institutions upholding the principles, then she is acting in ways consistent with political liberalism (Rawls 2005, pp.77-78). The source of her motivation to act in the way in which she does, whether it lies in an abstract understanding of the principles, conformity to the group, or an emotional commitment, is inconsequential. Rawls assumes that people will desire to be cooperative citizens, but he does not address the source of this desire (Rawls 2005, pp.81-88).

Nevertheless, a problem remains unresolved. Individuals being left to their own resources to develop a sense of morality may be motivated by influences at odds with justice as fairness. For instance, an individual's early experiences may have highlighted the importance of hard work and the individual will overcoming obstacles. This person may then come to hold a highly individualistic, politically libertarian understanding of fairness. On this understanding, it is fair if a person owns something if she worked for it, and it is unfair if any of this is taken away from her, even if it is taken to satisfy the basic needs of those less fortunate. People and communities existing within modern societies may hold such an understanding of fairness, which sits in tension with that of the Rawlsian understanding. While we might consider such people rationally autonomous, it is doubtful that they would act as fully autonomous citizens are expected to act. They would not be motivated to act from the principles of justice when making political decisions. Developing arguments presented in Chapter 3, I argue that if there is to be any hope of considering individuals as fully autonomous in the way that Rawls imagines, the development of autonomy must be the priority of education.

In Chapter 3, I argued against the position of James Scott Johnston (2005). Johnston argues that if education is to play any role within political liberalism, it should only to be to establish publicity; to make children understand the importance of their future role as citizens. However, given that empirical studies have shown that most people do not develop a sense of morality from which they act on abstract principles, but from emotional responses to what is perceived to be wrong or right (Haidt 2007), we cannot assume that people will become fully autonomous, or act from a common understanding of fairness, only from an understanding of their role. That is to say most people will not act from the principles of justice due to their being internally motivated to do so. Rather, people will accept the given rules of a society but without any deeper internal commitment to these rules. In the case that neural determinism is true, this would not be sufficient. Without an internal commitment to justice as fairness, citizens are vulnerable to the influence of unreasonable doctrines. To ensure the stability of political liberalism, fully autonomous citizens who are motivated to uphold the principles of justice are necessary. If education is aimed only at ensuring children understand the importance of citizenship, we cannot guarantee that citizens will develop the necessary capacities to be considered fully autonomous. Neither will their emotional commitments necessarily converge with the ideals of political liberalism. Without any commitment to the value of fairness within moral education, a person could develop either to disregard the importance of fairness, or think of fairness in a way entirely at odds with the Rawlsian conception.

Instead, if political liberalism is to work in the way that Johnston imagines, it would need to be a 'Government House' approach to liberalism (Williams 1985, pp.108-100). We cannot assume that the condition of publicity alone will lead to citizens being motivated to realise their full autonomy. In the case that many citizens do not, an elite of citizens who had developed to the final stage of moral development, acting in accord with abstract moral principles, could be considered fully autonomous, and create the rules by which the rest of the citizens, who are not necessarily fully autonomous, abide. As most people develop to the conventional level of moral development, obeying the rules established within the social order, this would be possible. Most citizens would only obey the principles of justice, however, there would be no deeper commitment to the principles. This 'Government House' approach to political liberalism reveals a flaw in Johnston's position. Such a system seems inherently unstable; as there is no motivation to endorse the principles of justice on the part of most citizens, ideas that threaten the stability of political liberalism could potentially gain support. Furthermore, there can be no guarantee the elites themselves would develop and maintain the appropriate sense of justice across time. Should politicians begin to question the ideals of political liberalism, citizens would have no internal motivation to continue supporting such ideals.

Rather than devise such a 'Government House' approach, I argue that the two moral powers should be developed through a system of moral education for all citizens, a system that is substantive and interventionist, ensuring children come to understand fairness in the Rawlsian sense. Children should understand both the principles of justice, and what it means to hold a reasonable conception of the good. Such an education system must be premised on an appreciation of the empirical data, however. Rather than assume all citizens will develop to understand and act from abstract principles, we should instead consider the emotional commitments people develop in childhood. If neural determinism is true, and we have little conscious control over our thoughts, then it cannot be guaranteed that citizens will develop full autonomy through their own capacities for independent thought. Fairness is of key importance here. We cannot assume that people will develop an understanding of the importance of fairness, or that there is a common understanding of what is fair. Thus, it should be ensured that people are encouraged to develop an emotional commitment to the value of fairness in the Rawlsian sense. From this commitment, we could better ensure that people come to recognise justice as fairness. This commitment is central to the development of fully autonomous citizens. Without such an approach to education within political liberalism, due to the implications of neural determinism, we could not guarantee that citizens would become fully autonomous.

As M. Victoria Costa notes (2004), however, teaching the importance of fairness would lead to the diversity of beliefs held in modern societies being lost. People educated to recognise the emotional significance of one moral value would be less inclined to endorse alternative doctrines. To establish the necessary sentiments to support justice as fairness, something akin to Cohen's "egalitarian ethos" would be necessary (Cohen 1992). The cost of establishing such an ethos would be the loss of diversity of beliefs held within a society, as Costa notes. I consider Cohen's explication of the egalitarian ethos, before moving to consider Titelbaum's rejoinder to Cohen (Titelbaum 2008). For the stability of a well-ordered society, I argue that something closer to Cohen's ethos is necessary. Full autonomy provides the telos for such an ethos. As Titelbaum recognises, however, establishing such an ethos will reduce the diversity of beliefs people hold.

Cohen argues that the egalitarian ethos is a prerequisite of a just society (Cohen 1992, pp.315-316). That is, people need to be motivated to act justly in order for a principle of justice to be upheld. In regard to the difference principle, Cohen views this as self-defeating. A society of just people who endorsed the difference principle due to their sense of justice would not be motivated by their self-interest to pursue high salaries. Thus, the inequalities the difference principle allows for would not obtain.

In his explanation of the "Rawlsian ethos of justice", Michael Titelbaum concedes that though something akin to G.A. Cohen's "egalitarian ethos" is not impossible to incorporate within the scheme of political liberalism, to allow for this would be to lose diversity at the expense of stability (Titelbaum 2008, pp.319-320). Titelbaum argues for what he calls the "full ethos" rather than the "egalitarian ethos" (Titelbaum 2008, p.306) The full ethos demands that people act from the principles of justice in their everyday lives, but does not go so far as to demand a commitment to egalitarianism across a person's life. Fostering this ethos would, however, reduce diversity of beliefs within a society. While the principles of justice place a certain limit on what is a permissible belief, by demanding that citizens act from the principles of justice in their everyday lives, the full ethos places further limits on what is permissible, though not to the same extent as the egalitarian ethos (Titelbaum 2008, pp.319-320). On the egalitarian ethos, a person is never justified in pursuing a higher salary, while on the full ethos, a person is justified in pursuing a higher salary if the reason for doing so is in accord with the principles of justice. Contra Titelbaum, I argue that a scheme of moral education that requires us to become fully autonomous would move us much closer to Cohen's egalitarian ethos. As Titelbaum notes, however, such an ethos comes at the price of diversity.

A person who had been educated to recognise the emotional significance of justice as fairness would have no motivation to develop a conception of the good in accordance with a reasonable doctrine – a doctrine that is reasonable on the grounds that it is compatible with other reasonable doctrines, and does not attempt to use state power to suppress such doctrines – that was at odds with justice as fairness in certain ways. This is particularly of concern for the diversity of political beliefs. For instance, it seems unlikely a person who had been raised to appreciate fairness in the Rawlsian sense would develop a commitment to political libertarianism. As Samuel Scheffler notes (1994), there is a contradiction between the way in which political liberalism seeks to protect diversity of reasonable beliefs while also promoting a certain conception of justice. While this is a concern for the diversity of political beliefs held in a society, the motivation to endorse non-political doctrines will also be reduced. A person who had been educated in the Rawlsian value of fairness, and who lived a life in accord with this value, would have little need of the moral teachings found in religious doctrines.

Furthermore, if people have little conscious control over their thoughts, which are determined beyond their control, and it is our emotional commitments which determine our sense of morality, rather than our capacity for reason, if a person did develop a conception of the good drawn from libertarianism, she may not be able to ensure her sense of justice stayed in accord with justice as fairness. Hence, there are two problems here. First, the problem of motivation, and second, the problem of lack of control. We must ensure that people are motivated to endorse the principles of justice, as this supports the stability of political liberalism. As people have little conscious control over their thoughts, we should also ensure that people are raised in environments where their experiences enable such a motivation. By taking such steps, however, we reduce the likelihood of diversity within the senses of morality which people hold.

Divergence between the right and the good is what protects the diversity of beliefs. While a person under political liberalism must develop their first moral power so as to recognise justice as fairness, she is free to develop her second moral power, her ability to form a conception of the good, as she wishes, providing the way in which she conceives of the good is within the bounds of reasonableness. Whereas Rawls argues that the right and the good are congruent in *A Theory of Justice* – that it is to be considered good to act justly (Rawls 1971, pp.567-572), and that we come to our conception of the good within the limits of what is right (Rawls 1971, pp.563-564) – in *Political Liberalism*, arguments for congruence are dropped. As Samuel Freeman puts it, though Rawls never rejected his own belief in the truth of justice as fairness, that a doctrine is "*philosophically* justifiable" is not to say that it is "*publicly justifiable"* to the members of a democratic society (Freeman 2003, p.325). Jon Mandle agrees with Freeman here, adding that the task of political liberalism is to make justice as fairness justifiable to reasonable people, with the answer being found in the idea of an overlapping consensus (Mandle 2009, pp.22-23). However, these reasonable people may not necessarily view acting from the principles of justice as a good in itself. People may have conflicting moral motivations within political liberalism.

This, I argue, presents a third problem. If, empirically speaking, most people struggle to realise a sense of morality from which they can understand and act from abstract moral principles, it would appear to be unlikely that a person could differentiate between the right and the good, understanding the moral doctrines that lead to the formulation of either. As Kohlberg eventually recognised, not everyone can be a moral philosopher (Gibbs 2013, p.90). A person raised with a strong religious faith is unlikely to understand how her sense of the good diverges from her sense of the right, unless she has philosophical training. Should the two come into conflict, it could be assumed that she would not understand the source of the conflict. If Haidt's theory of moral development is correct (2007), and we do come to conclusions as to what is moral based on feeling rather than reason, then we should ensure that the scheme of moral education a person develops through focuses on fairness, and that the individual is emotionally committed to the importance of the Rawlsian understanding of fairness. Of course, this is conditional on the truth of these psychological claims in regard to both the work of Kohlberg and Haidt. Nevertheless, even in the absence of such claims, if most citizens are assumed not to be moral philosophers, it would be appropriate to expect citizens to live according to unified moral motivations. The conflict of moral motivations should thus be reduced if we are to prioritise the development of fully autonomous citizens.

To conclude, there are three problems with education for political liberalism. First, citizens will not necessarily be motivated to endorse justice as fairness. Citizens whose experiences have led them to hold a sense of morality that does not cohere with justice as fairness are unlikely to become fully autonomous. Second, citizens have little ability to consciously control their thoughts. We cannot assume that where such conflicts arise, a citizen will be able to regulate her thoughts in accord with the values of justice as fairness. Third, political liberalism allows for the conflict of moral motivations. A person's sense of the good may be drawn from a different moral tradition than her sense of the right. Unless a society consists entirely of moral philosophers, most people will not understand the conflict between competing moral motivations. The source of this conflict should, then, be reduced through moral education. If a well-ordered society is to be devised, citizens must be encouraged to become fully autonomous through their system of education. The focus of this scheme of education must be perfectionist. Full autonomy provides the standard of the good life which it is then the aim of the state to promote. It is in this way that political liberalism can incorporate certain perfectionist measures. Through employing such measures, stability is thus enhanced, though the risk here is that diversity is lost.

9.5 Perfectionist Education

As I have argued, the focus of moral education should be perfectionist in character. In this section, I further explain this scheme of education. Returning to the examples of David, Syeda, and Abdul, I argue that such people would be less likely to act in unreasonable ways if the focus of moral education was perfectionist. The aim of this perfectionist moral education should be to foster an ethos from which citizens come to realise the importance of full autonomy. Acting from such an ethos, citizens would not act in ways that erode the sense of fairness within society. A perfectionist scheme of education would instil a motivation in individuals to become fully autonomous, reducing the need for paternalistic acts against adults.

Before proceeding, it should be noted that all of what follows is considered within the context of non-ideal theory. It is not supposed that a state could be reached in which every single inhabitant of a society develops to be a fully autonomous citizen. Nor is it thought that all beliefs contrary to justice as fairness would necessarily disappear from society. Complete uniformity of belief would be neither possible nor desirable. The aim of this section is rather to establish the conditions necessary under which we can consider individuals as being able to develop to be fully autonomous citizens.

To return to the example of David, we can imagine that David's adherence to monetarism was derived from his conception of the good, which is drawn from the doctrine of political libertarianism. Thus, David's conception of the good is expressed through the importance of negative freedom, individualism, and independence, while his sense of justice should be characterised by the importance of fairness. Perhaps if David is one of the few people who develops to the final stage of moral development, and acts from an understanding of abstract moral principles, then such conflicting moral motivations will not pose a problem. However, as we cannot assume that all people are capable of such an understanding of morality, the sense of morality encouraged within a society should be simplified. Rather than expecting a person to develop a complex moral psychology, through which she could follow one system of thought in regard to her understanding of the right, and another in her understanding of the good, through moral education, her moral motivations should become unified. This is not to say that all understandings of the good across society need necessarily to be unified; it is perfectly possible that some people will develop different understandings of the good that are entirely reasonable, and in no way hinder the development of their full autonomy. Instead, it is to say that if we are to allow for the development of rationally and fully autonomous citizens, divergence should be discouraged, particularly in regard to education for children. The importance of fairness must be the primary moral motivation, guiding the development of both moral powers.

From considering the implications of neural determinism, we could not imagine that David would consciously control his thoughts in order to ensure his conception of the good did not influence his journalistic writings. A person committed to a doctrine is likely to be influenced by that doctrine across her life, and given the emotional commitments people have towards the doctrines which they endorse, people are unlikely to compartmentalise these commitments. Rather than attempt to censor David's writings, we should instead aim for a system of education under which it would be less likely that someone would emerge convinced by libertarianism. To avoid paternalistic acts against David as an adult, we should inculcate a desire to act in ways that will uphold the idea of a well-ordered society when David is a child. Thus, David's motivation to endorse libertarianism – or other doctrines that, though reasonable, could act against the principles of justice – as an adult should be reduced. To avoid paternalism against adults, a scheme of perfectionist education for children is necessary, where the aim is the development of full autonomy.

Ben Colburn has argued that education that aims to develop autonomy cannot be considered perfectionist (Colburn 2008, pp.623-624). This is because promoting autonomy does not necessarily mean people have to be committed to the value of autonomy. It could be that people live autonomous lives without considering autonomy as an important value. An education system that promotes autonomy does not, then, have to instil a sense of the importance of autonomy, as people do not have to be aware of autonomy's importance to live autonomous lives. There is a difference, however, between the type of autonomy Colburn aims to promote and that which is under discussion here. As Colburn notes, the second moral power as defined by Rawls is synonymous with the type of autonomy Colburn wishes to promote (Colburn 2010, p.66). A person who has the capacity to choose the good for herself and live a life in accordance possesses the second moral power, on Rawls' view, and is autonomous, on Colburn's view. As noted in Chapter 3, while rational autonomy is similar to the type of autonomy promoted by Raz and Colburn, full autonomy is closer to the Kantian conception of autonomy. An education system that aims to promote full autonomy would necessarily be perfectionist. The fully autonomous citizen is not one who chooses the good for herself, living a life in accord, but one who lives her life justly. Promoting such a way of life necessarily presupposes a prior standard of the good, meaning an education system premised on promoting full autonomy could not be considered non-perfectionist. The aim of this type of education is to encourage people to act justly so as to reduce the need to act paternalistically against adults.

The assumption here is that what would be considered a paternalistic act against an adult is not paternalistic, or at least not unjustifiably paternalistic, if committed against a child. John Stuart Mill argued that this was the case (Mill 1859), while liberals since Mill have generally agreed. Dworkin agrees with Mill that arguments against paternalistic restrictions of freedom only apply to "mature individuals" (Dworkin 1972). Quong only applies his argument against paternalism to "sane adult citizens" (Quong 2011, p.86). Amy Gutmann notes a problem with acting paternalistically against children: as justificatory arguments for paternalism often rest on consent, which children as non-rational agents cannot give, paternalism against children cannot be justified (Gutmann 1980, pp.339-340). Nevertheless, Gutmann concludes that paternalism can be justified against children on the Rawlsian grounds of the need for primary goods. As there are certain goods rational adults would have wanted to have as children – nutrition, healthcare, housing, education – then paternalistic acts which aim to provide such goods to children are justified. If a child was to reject such goods, paternalism against the child would not be unjustified. Rawls, as aforementioned, is also not concerned with paternalism in regard to irrational agents, where their irrationality leads them to reject primary goods or their more specific aims in life (Rawls 1971, pp.248-250). Education is bound to respect such constraints on freedom, too. Though Rawls does not explicitly offer an explanation as to why this is the case, it can be assumed that it is because we cannot assume the rationality of children. Where children are liable to act in irrational ways, paternalism is justified against them.

I agree that the arguments of Gutmann and Rawls are correct. Overlooked in these arguments, however, are the moral powers. It is not only that paternalism can be justified in regard to children on the grounds of ensuring access to primary goods, but also to develop a child's sense of morality. This, for Mill, was part of the purpose of education (1859). Because of the implications of neural determinism, it cannot be assumed that people left to their own devices will become fully autonomous. Thus, we cannot assume that an individual has a sense of morality that is inherently bound to that necessary for justice as fairness. Therefore, an individual's sense of morality must be the focus of education. As the child's sense of morality is yet to fully develop, however, paternalistic acts against her cannot be considered unjustified, at least on Quong's argument against paternalism. If a child does not possess the second moral power, acts against her cannot contravene her ability to determine her own good, as she has no such ability. Thus, whether or not such acts against children are considered to be paternalistic, they are not unjustified.

Perfectionist education is therefore justified; it cannot be considered unjust due to its being paternalistic. The aim of such a scheme is to inculcate a disposition in the individual towards full autonomy. It does so through establishing an emotional commitment to the importance of fairness in the Rawlsian sense. With such a scheme established, individuals are motivated to act as fully autonomous citizens due to their shared understanding of the importance of fairness. An ethos not unlike that of Cohen's egalitarian ethos is thus secured, with citizens being guided by such an ethos in their acts. As Cohen, following Marx, notes, in the case of the appropriate ethos being established, the "state can wither away" (Marx 1843, p.241; Cohen 2008, pp.1-2). Whether or not we follow Marx in imagining that the state would necessarily begin to dissolve in the case of citizens following the appropriate ethos, with citizens being motivated to act in accord with the principles of justice, there will be less need for the use of state power. We are, then, perfectionistic and paternalistic towards children so as not to be paternalistic against adults.

Following such a perfectionist scheme of education, David, Abdul, and Syeda would be motivated to develop their full autonomy. David's journalism would argue in favour of economic principles that support justice as fairness. Syeda would think more about the way in which people may be influenced by ideas expressed on her social media website. Furthermore, users of her website would be less inclined to express unreasonable ideas with a perfectionist scheme of education in place. Abdul would not market a product that led children to develop eating disorders, nor would his employers wish him to do so. Though paternalism would not necessarily be unjustified against David, Syeda, or Abdul if they developed their moral powers in ways inconsistent with justice as fairness, rather than act paternalistically against an adult, a perfectionist scheme of education would lead individuals to develop their moral powers in the appropriate way. The result of this would be a reduction in the diversity of beliefs within a society – someone such as David would be less likely to endorse libertarianism, for instance – but the stability of society would be enhanced.

To conclude, if we are to ensure the development of rationally and fully autonomous citizens, we must ensure a system of perfectionist education. Citizens as adults must be motivated to endorse and act from the principles of justice. Therefore, we cannot assume this motivation is inherent in the individual; instead, we should ensure it exists in children through the system of education. Through education, children should become motivated to endorse justice as fairness. Reasonable doctrines that are nevertheless in some ways at odds with justice as fairness should be discouraged. The effect of this is likely to reduce the diversity of beliefs within a society, yet to also ensure citizens are motivated to uphold the idea of a well-ordered society based on terms of fair cooperation.

9.6 Conclusion

A consideration of the data within neuroscience should lead us to recognise the vulnerability of autonomy. Without the existence of the autonomous citizen, the stability of political liberalism is lost. Thus, the development of autonomy should be prioritised. In this chapter, I have explored the consequences of such a move. The central problem is the unreasonable person. Against Quong, I have claimed that not all unreasonable people are Nazis, or similar extremists. Many people act in ways which undermine the terms of fair cooperation and limit the development of autonomy for other people. If our task is to develop a society in which autonomy can flourish, then we need to consider the acts of people who impede the development of autonomy for others. The task of persuading unreasonable people to endorse political liberalism is not one of persuading a minority of citizens who hold extreme views, but of changing the behaviour of many people across society.

To endorse such a move is to be left open to the accusation of paternalism. While I argue we should not be unconcerned about this criticism, I claim that concerns regarding paternalism are exaggerated. Paternalism is sometimes necessary, and it is not the case that the political liberal has no standard of the good to measure against. Where acts undermine the sense of fairness, and erode the ability of people to become rationally and fully autonomous, as such acts are unreasonable or irrational, we are justified in acting paternalistically to prevent such acts. This is not to say we should desire a paternalistic state. However, if we are to maintain a society of autonomous citizens, then certain paternalistic acts will be necessary. Rather than act paternalistically towards adults, it is preferable to ensure the development of autonomy through the education of children. On this view, the aim of education is to promote the value of full autonomy, reducing the motivation of people to think unreasonable ideas, or to act in unreasonable ways. Thus, our scheme of moral education should be perfectionist. With such a scheme in place, the problem of unreasonableness should fade over time.

Chapter 10 – Conclusion

10.1 Perfectionism and Full Autonomy

If we are to realise the well-ordered society Rawls imagines, and we take the problem of neural determinism seriously, perfectionist measures are needed within political liberalism. People cannot be relied on to become fully autonomous – acting from principles of justice – drawing only on their own capacities to choose for themselves. The way in which they are influenced by ideas is beyond their control. In a society containing many unreasonable ideas, citizens cannot be expected to become fully autonomous if they come to endorse such unreasonable ideas. Instead, full autonomy must be viewed as a perfectionist ideal that citizens are expected to live in accord with. Thus, there is a standard of the good within political liberalism against which we can judge. A citizen who is fully autonomous and committed to acting justly would not be influenced by unreasonable doctrines, nor would she seek to influence others through such doctrines. No sense of justice within a society could thrive without people concerned to uphold it. Thus, if the Rawlsian theory of justice is to be maintained within a society, there must be a scheme of moral education that motivates people to endorse justice as fairness.

I begin by restating my basic argument. Following this, I assess the desirability of this approach to political theory. The liberal who views liberalism as founded on the importance of individual freedom may be repelled by the idea of imposing a moral doctrine on children. Against this, I argue that if we are to protect liberal values such as freedom of choice, we need citizens who are motivated to uphold these values. In a society of fully autonomous citizens, it would be unnecessary to act paternalistically. Thus, this revision of political liberalism is not authoritarian or illiberal. Finally, I examine some of the practical ramifications of this approach. I argue that approaching political liberalism in this way would alter the policy preferences citizens hold.

10.2 The Perfectionism Necessary within Political Liberalism

As Rawls recognises that citizens will often disagree with one another in a democracy, the task of his political theory is to identify a constitutional form that will allow for the reconciliation of these disagreements (Rawls 1971, pp.195-201). The constitution he imagines is undergirded by the two principles of justice. Equal rights and basic liberties guaranteed by these principles are thus enshrined in the constitution. With this established, citizens have a duty to act in accord with the constitution, but do not have obligations beyond this. If we imagine that science is considered within political liberalism – at all levels of enquiry, from the original position to political deliberation between citizens – though there would not be a revision of what are considered essentials of the constitution, it would be recognised that citizens have additional duties beyond obeying the rules of the constitution. A deeper commitment is needed to the principles of justice, and citizens would need to ensure this commitment is established.

For Rawls, the two principles of justice establish how the constitution should be arranged. That is, each person having a claim to the same rights and liberties as all others, and social and economic inequalities are arranged to the benefit of the least advantaged (Rawls 2005, pp.5-6). Parties in the original position would recognise that these principles can only obtain if there are people motivated to uphold them. If the truth of Rawls' account of moral psychology is accepted, then it can be assumed that people will be naturally inclined to accept these principles. However, for reasons I have argued throughout this thesis, we should not accept this truth. Instead, we should assess the empirical evidence within psychology and neuroscience. On considering this within the original position, parties would recognise that we cannot rely on a person accepting the principles of justice due to her innate moral psychology, or her capacity for free will. If the principles of justice are to be accepted, society should be devised so as to foster the appropriate form of moral psychology. This requires more than ensuring the just workings of institutions. Even in a society with perfectly just institutions, a person may not be committed to the principles of justice if her moral psychology is not aligned with the Rawlsian sense of fairness.

Acting against an adult to ensure her moral psychology cohered with this sense of fairness would undermine the liberty of conscience enshrined in the first principle of justice, meaning this would become self-defeating. As Amy Guttman has argued (1980), paternalism against children can be justified on Rawlsian grounds; we should, then, be perfectionist and paternalistic towards children to avoid such acts of paternalism against adults. Through such perfectionist measures, the development of moral psychology could be fostered through education. The aim of such education is the development of fully autonomous citizens motivated to uphold the principles of justice.

When determining how the constitution should be arranged in accordance with the principles of justice, these implications should be considered. If the stability of the constitution is to be realised, fully autonomous citizens need to uphold the principles of justice. Under principles that guaranteed equal rights and liberties, and arranged social inequalities to the benefit of the least advantaged, it is still possible that a person would use these rights and liberties to the detriment of these principles. Therefore, if a just and stable constitution is to be supported within the kind of well-ordered society imagined by Rawls, a commitment to the development of full autonomy must be established. Through being perfectionist in regard to children, we avoid the need to be paternalistic against adults in the task of ensuring the development of full autonomy. With this established, the stability of a liberal political order is better ensured.

10.3 Is this Desirable?

It may be thought that this is an illiberal approach, and therefore undesirable. One might take into consideration the implications of neuroscience, but still maintain a commitment to the importance of freedom of choice. Regardless of the level of control we hold over our thoughts, a person may believe that the protection of freedom of choice is sacrosanct. If a person chooses to endorse political libertarianism, as in the example of David in Chapter 9, this should be his choice to make, as the author of his own life. This bears on what a person views as being the aim of liberalism. One might view rational autonomy as being of greater importance than full autonomy. If the aim of liberalism is to enable people to lead a life of their own choosing, and all the other ends of liberalism are subordinate to the protection of individual choice, then it is irreconcilable with this revision of political liberalism. Instead, this revision prioritises the stability of a well-ordered society supported by a just constitution. Within this, people will still be able to live lives of their own choosing, but choice will be subordinate to stability. I argue that if the end of liberalism is to enable freedom of choice – as in the tradition of Mill's approach to liberalism – then this revision of political liberalism will not be appealing. However, I explain why this conception of liberalism should not necessarily be viewed as unattractive or illiberal.

Jean Jacques Rousseau argued that if a person refused to obey the general will, he would be forced to do so (Rousseau 1762). In this sense, Rousseau thought the person was being "forced to be free". By taking an approach to political liberalism that prioritises full autonomy, I am not arguing that a person must be forced to be fully autonomous. Instead, the formulation of society should be such that a person desires to become fully autonomous. Rather than being *forced* to be autonomous, my argument is that people must be *encouraged* to consider the importance of fairness. There is no argument here for the power of the state to be used to force people to act justly across their lives. It is instead a matter of the culture of a society being such that a person is motivated to act as a fully autonomous citizen. As argued in Chapter 9, perfectionist measures are taken in regard to children to avoid paternalism against adults. Rather than interfering with the freedom of adults, we should encourage the appropriate dispositions in children so that, as adults, people are motivated to act justly.

The liberal who views liberalism as promoting choice, enabling people to lead lives of their own choosing, may be troubled by the revision of political liberalism I have offered. Liberals in the tradition of John Stuart Mill (1859) may be particular troubled by this. For instance, Joseph Raz, who views liberalism as founded on the promotion of individual liberty (Raz 1986, p.367), would presumably not desire a state that prioritised the promotion of a particular way of life over the protection of individual freedom. If choice is central to what is important in our lives, and the state should protect our ability to make choices, it would seem contradictory for the state to impose a certain sense of morality on people through its education system. As it is for people to develop their second moral power, a person’s moral commitments are for her to determine. On this revision of political liberalism, however, I am not arguing that choice needs to be prevented. Instead, we should ensure that people have the necessary dispositions to make good choices as adults. It should not be assumed that people are born with such dispositions. If a liberal society with a just constitution is to be sustainable, a citizenry with the appropriate moral psychology is necessary. The development of such a moral psychology should be prioritised over the protection of individual free choice. Once a person has reached adulthood, however, there is no argument for the use of state power in relation to her moral choices. As argued in Chapter 9, perfectionism is utilised for children to avoid being paternalistic towards adults.

However, this is not to assume that society will be perfectly just, or that all citizens will necessarily accept justice as fairness. There is nothing to prevent citizens from choosing to endorse alternative doctrines. If as an adult, a person chooses to endorse political libertarianism, this choice should be respected; the power of the state should not be used in an attempt to prevent this choice, nor should the person be sanctioned This person would pose little threat to stability in a society largely committed to justice as fairness. No use of state power would be used to prevent people from endorsing the beliefs they wish to believe. In a society committed to justice as fairness, the first principle of justice would protect freedom of conscience; if "each person has an equal claim to a fully adequate scheme of equal basic rights and liberties" (Rawls 2005, p.5), the state could not infringe one person's right to freedom of conscience without imposing the same infringement on all others. To do so would be self-defeating. This approach to liberalism is premised on the need to develop a society which is committed to liberal ideals. It cannot be assumed that people will necessarily come to endorse liberal ideals through their own capacity for free will, or an innate sense of morality. This is why a perfectionist scheme of moral education is necessary. If the appropriate moral sentiments that enable a person to become fully autonomous are developed in children, it is not necessary to act paternalistically against adults. Once people have developed to be fully autonomous, there is no need for the use of state power in matters pertaining to individual morality; it can be assumed that most of the choices a person makes will be just if she has a commitment to the principles of justice.

In this sense, individual choices can be respected with less concern regarding the consequences of these choices than would be the case in a society prioritising the protection of individual choice. In the latter case, people may use the freedoms granted to them to choose to vote for policies which would end those same freedoms. With no prior commitment to the importance of these freedoms on the part of citizens, there is little to prevent citizens from acting in this way, if they are persuaded to do so. However, citizens with a prior commitment to the importance of full autonomy and acting justly are unlikely to be persuaded by unreasonable ideas. Thus, through promoting the development of full autonomy, freedom of choice is better protected than under a scheme which primarily aims to protect freedom of choice.

Hence, this revision of political liberalism should not necessarily be viewed as unattractively authoritarian or illiberal. It is instead an argument to protect liberal values in light of the problems revealed through the empirical evidence. Rather than assume people will act as citizens should in a liberal democracy providing the institutions are just, society should be structured as to ensure fully autonomous citizens who are motivated to act justly. The proponent of political liberalism who holds that constitutional stability is to be attained should take the implications of neural determinism seriously. If the aim of liberalism is a society of reasonable people upholding a just constitution, the development of moral psychology needs to be considered. To reach a state in which the majority of citizens are reasonable and act in accord with the prescriptions of full autonomy, a scheme of moral education would be needed, the consequence of which would be a reduction in the diversity of beliefs within society. Nevertheless, the liberal who, above all else, wishes to protect freedom of choice and this diversity of beliefs will no doubt not be persuaded by this formulation of liberalism.

10.4 Beyond the Constitution

Having explained why I view this revision of political liberalism as being more stable, I now turn to how it may help us respond to other political, social, and environmental issues. People viewing themselves as autonomous, but recognising the limits to their capacity to exercise free choice, would favour different political principles from those they would endorse on Rawls' formulation of political liberalism. As aforementioned, Rawls was primarily concerned with the basic structure of society. Thus, his theory of justice aims to identify the type of constitution which will support a just society through just institutions. On this revision of political liberalism, citizens will favour a slightly altered constitution, but also recognise that the attainment of justice requires other citizens motivated to act justly. In this section, I assess possible adjustments to the constitution, and how the political behaviour of citizens might be altered.

Spinoza, explaining the ethical consequences that followed from his metaphysical doctrine – according to which, we are all part of the divine nature, and everything that exists does so out of necessity – argues that, on accepting this doctrine, we would change our social lives. This doctrine, Spinoza says:

> Teaches us to hate no one, to disesteem no one, to mock no one, to be angry at no one, to envy no one (Spinoza 1677, p.68).

Spinoza goes on to claim that, with our reason reflecting on this doctrine, we would be compelled to help each other. Likewise, J.J.C. Smart, in an essay critiquing the libertarian position on free will (Smart 1961), argued that if determinism was accepted as true, we should be less inclined to judge and blame people for their failings. I argue that the fully autonomous citizen, recognising the importance of acting justly, but also the vulnerability of her own autonomy – that her autonomy is the result of the social environment in which she lives – would accept the propositions offered by Spinoza and Smart. While the fully autonomous citizen is motivated to act justly, she does not castigate those who fail in their moral duties. Instead, she recognises the need to structure society so as to increase the likelihood of people acting justly.

Thus, on this approach to political liberalism, fully autonomous citizens would alter their political behaviour. Fundamentally, on this approach, less emphasis is placed on the importance of free choice, and more on the need to act justly. If we cannot rely on a person to develop the appropriate sense of justice through her own capacity for free will, this motivation must be encouraged in her from childhood. I now offer brief explanations of how this approach may alter political and social behaviour.

A different attitude may be taken towards the kind of inequalities existing within a society. As Cohen argued, the kinds of inequalities considered just on the difference principle would not obtain in a society in which there was an egalitarian ethos (Cohen 2008). Similarly, in a society of fully autonomous citizens, people would not act in ways that would undermine the ability of others to realise their own rational and full autonomy. A successful entrepreneur would not seek to protect her wealth at the expense of the least advantaged. Furthermore, as citizens recognise the limited capacity people have to consciously control their thought processes, less blame would be placed on people for economic failures. If people can do little to prevent their thought processes being influenced by phenomena they encounter, then there may be little they can do to prevent bad economic choices as a result of these influences. Thus, a different attitude toward welfare policies could be justified on this basis, one that is more egalitarian and less acquiescent towards inequalities.

As in the example of Abdul in Chapter 9, fully autonomous citizens would not act in ways that would undermine the rational autonomy of other citizens. An adult whose childhood addictions had stymied her ability to determine a rational plan of life could not be rationally autonomous. Thus, a fully autonomous citizen would not engage in activities which could affect others in this way. She would not consider it the duty of other citizens to exercise their capacity to choose freely in order to overcome such obstacles. Instead, acting justly, she would not engage in such practices. Thus, in a society of fully autonomous citizens, businesses would not desire to sell harmful products, journalists would not deceive their readers, nor would gambling venues allow people to lose their life savings. It would be recognised that such acts prevent others from developing their own autonomy.

Finally, a fully autonomous citizen, considering the implications of neural determinism, would be less likely to desire a punitive justice system. Being less inclined to blame and judge others for their failings, fully autonomous citizens would think instead about how the structure of society influences the choices a person makes. Reducing the causes of crime and focusing on rehabilitation would be prioritised over retribution. Fundamentally, the aim would be to instil in people the importance of acting justly that would reduce the likelihood of people acting against the rules of fair cooperation. Thus, focusing on the development of full autonomy and the sense of morality it promotes would settle certain questions of law and order.

That such changes to society would be made, along with the ways of life within society, gives us further reason to consider the implications of neuroscience. On considering these implications, and realising the importance of full autonomy, citizens would alter their political preferences. Instilling a sense of the importance of fairness in people is to prepare them for life in a cooperative society supported by a just constitution. Citizens recognising the importance of fairness would be committed to upholding the ideals that maintain such a society and constitution. However, this requires more than recognition and understanding of how the basic structure of society upholds just institutions. It requires citizens to be committed to the values of political liberalism and to act justly across their lives.

Glossary

Action Potential – The state of a neurone when its voltage is increased as it communicates with other neurones.

Amygdala – Part of the limbic system concerned with emotion.

Anterior – At the front.

Basal Ganglia – A group of large nuclei that regulate both voluntary and involuntary movements.

Cerebellum – A structure at the back of the brain that allows for conscious awareness of movement and the position of the body.

Cerebral Cortex – The main section of the cerebrum, and the section responsible for a large part of our thought processes.

Cingulate Cortex – Part of the cerebral cortex that regulates emotional responses in conjunction with the amygdala in the limbic system.

Dorsal – At the back or top.

Lateral – At the side.

Limbic System – The name for a system including several parts of the brain – the amygdala, hippocampus, limbic cortex, and septal area – responsible for memory and emotion.

Motor Cortex – The control centre for our bodily movements, including the primary motor cortex, pre-motor cortex, and the supplementary motor area.

Nucleus (plural: nuclei) – A group of neurones.

Neurone – The cells contained in the brain and nervous system that allow for the communication of information throughout the brain and body.

Posterior – At the back.

Posterior Superior Temporal Sulcus – Contained within Wernicke's area, a part concerned with speech, along with interpreting the intentions of others.

Prefrontal Cortex – Part of the cerebral cortex involved in the process of memory and learning.

Pre-Motor Cortex – Part of the motor cortex responsible for modulating actions.

Primary Motor Cortex – The main part of the motor cortex, and a part playing a large role in voluntary movements.

Readiness Potential – Activity in the brain occurring before the body acts.

Striatum – A nucleus that transmits dopamine found within the basal ganglia.

Substantia Nigra – Part of the basal ganglia and transmits dopamine via the striatum.

Supplementary Motor Area – Part of the motor cortex that regulates body movement.

Thalamus – A pair of structures that allow for the conveying of sensory information.

Ventral – At the bottom.

Wernicke's Area – Plays an important role in speech and found within the cerebral cortex.

List of References

Amodio, D.M. (2014). The Neuroscience of Prejudice and Stereotyping. *Nature Reviews Neuroscience, 15*(10), pp.670-682.

Anderson, E. (2011). Democracy, Public Policy, and Lay Assessments of Scientific Testimony. *Episteme, 8*(2), pp.144-164.

Anderson, J. and Honneth, A. (2005). Autonomy, Vulnerability, Recognition, and Justice. In: Christman, J. and Anderson, J., ed., *Autonomy and the Challenges to Liberalism: New Essays*. Cambridge University Press, pp.127-149.

Aquinas, T. (1947). *Summa Theologica, I*. Translated by Fathers of the English Dominican Province. London: Burns & Oates.

Arendt, H. (1963). *Eichmann in Jerusalem: A Report on the Banality of Evil*. London: Penguin Classics.

Arneson, R. (2008). Rawls, Responsibility, and Distributive Justice. In: Fleurbaey, M., Salles, M., Weymark, J.A., ed., *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*. Cambridge: Cambridge University Press.

Arrow, K. (2003). Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice. In: Kukathas, C. Ed., *John Rawls: Critical Assessments of Leading Political Philosophers: Volume 1 – Foundations and Methods*. London: Routledge.

Augustine (1887). Rebuke and Grace. *Theological Anthropology*, pp.96-108.

Augustine (1950). *The City of God*. Translated by Dods, M.. New York: Modern Library, 2000.

Bach, D.R. and Dayan, P. (2017). Algorithms for Survival: A Comparative Perspective on Emotions. *Nature Reviews Neuroscience, 18*(5), pp.311-319.

Baldwin, T. (2008). Rawls and Moral Psychology. *Oxford Studies in Metaethics, 3*, pp.247-271.

Baumann, N. & Kuhl, J. (2003). Self-infiltration: Confusing Assigned Tasks as Self-Selected in Memory. *Personality and Social Psychology Bulletin, 29*(4), pp.487-497.
Andersen, F. and Becker Arenhart, J.R. (2016). Metaphysics within Science: Against Radical Naturalism. *Metaphilosophy*, *47*(2), pp.159-180.

Berlin, I. (1990). *The Crooked Timber of Humanity: Chapters in the History of Ideas – Second Edition*. Hardy, H., ed.. London: Pimlico, 2013.

Berlin, I. (1996). *The Sense of Reality: Studies in Ideas and their History*. Hardy, H., ed., London: Pimlico.

Böcker, K.B., Brunia, C.H. and Cluitmans, P.J. (1994). A Spatio-Temporal Dipole Model of the Readiness Potential in Humans. I. Finger Movement. *Electroencephalography and Clinical Neurophysiology, 91*(4), pp.275-285.

Bode, S., Murawski, C., Soon, C.S., Bode, P., Stahl, J. and Smith, P.L. (2014). Demystifying "Free Will": The Role of Contextual Information and Evidence Accumulation for Predictive Brain Activity. *Neuroscience & Biobehavioral Reviews*, 47, pp.636-645.

Boyd, D.R. (2015). *Becoming of Two Minds about Liberalism: A Chronicle of Philosophical and Moral Development*. Springer.

Bramhall, J. Discourse of Liberty and Necessity. In: Chappell, V., ed., *Hobbes and Bramhall on Liberty and Necessity*, 1999. Cambridge: Cambridge University Press.

Brass, M. and Haggard, P. (2007). To Do or Not to Do: The Neural Signature of Self-Control. *Journal of Neuroscience, 27*(34), pp.9141-9145.

Brembs, B. (2010). Towards a Scientific Concept of Free Will as a Biological Trait: Spontaneous Actions and Decision-Making in Invertebrates. *Proceedings of the Royal Society of London, 15th December 2010. Biological Sciences*. Available at:

http://rspb.royalsocietypublishing.org/content/early/2010/12/14/rspb.2010.2325.short (Accessed 26.10.2018).

Bush, G., Luu, P. and Posner, M.I. (2000). Cognitive and Emotional Influences in Anterior Cingulate Cortex. *Trends in Cognitive Sciences, 4*(6), pp.215-222.

Cacioppo, J.T. and Berntson, G.G. (2012). Is Consciousness Epiphenomenal? Social Neuroscience and the Case for Interacting Brains. *Social Neuroscience, 3*, pp.31-50.

Callan, E. (1996). Political Liberalism and Political Education. *The Review of Politics, 58*(1), pp.5-33.

Callender, C. (2011). Philosophy of Science and Metaphysics. In: *The Continuum Companion to the Philosophy of Science*, pp.33-54.

Carnap, R. (1963). Intellectual Autobriography. In: Schilpp, P.A., *The Philosophy of Rudolf Carnap*. La Salle: Open Court.

Charlemagne (2020). How Hungary's Leader, Viktor Orban, Gets Away With It. *The Economist*, 2nd April.

Churchland, P. (2011). *Braintrust: What Neuroscience Tells us About Morality*. Princeton: Princeton University Press.

Cohen, G.A. (1992). Incentives, inequality, and community. *The Tanner Lectures on Human Values, 13*, pp.263-329.

Cohen, G.A. (2008). *Rescuing Justice and Equality*. London: Harvard University Press.

Colburn, B. (2008). Forbidden Ways of Life. *The Philosophical Quarterly, 58*(233), pp.618-629.

Colburn, B. (2010). *Autonomy and Liberalism*. Abingdon: Routledge.

Colby, A., Kohlberg, L., Gibbs, J., Lieberman, M., Fischer, K. and Saltzstein, H.D. (1983). A Longitudinal Study of Moral Judgment. *Monographs of the Society for Research in Child Development*, pp.1-124.

Costa, M.V. (2004). Rawlsian Civic Education: Political not Minimal. *Journal of Applied Philosophy, 21*(1), pp.1-14.

Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy, 76*(5), pp.256-282.

Deecke, L., Scheid, P. and Kornhuber, H.H. (1969). Distribution of Readiness Potential, Pre-Motion Positivity, and Motor Potential of the Human Cerebral Cortex Preceding Voluntary Finger Movements. *Experimental Brain Research, 7*(2), pp.158-168.

De Ruiter, J.P., Noordzij, M., Newman-Norlund, S., Hagoort, P. & Toni, I. (2007). On the Origin of Intentions. *Attention & Performance XXII*, pp.593-610.

Dennett, D.C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: MIT Press.

Dennett, D.C. (2004). *Freedom Evolves*. London: Penguin Books.

Dreben, B. (2003). On Rawls and Political Liberalism. In: Freeman, S., ed*., The Cambridge Companion to Rawls*, Cambridge: Cambridge University Press.

Dubljević, V. (2013). Autonomy in Neuroethics: Political and not Metaphysical. *AJOB Neuroscience, 4*(4), pp.44-51.

Durkheim, E. (2013). *Durkheim: The Rules of Sociological Method: And Selected Texts on Sociology and its Method*. Edited by Lukes, S. New York: Palgrave Macmillan.

Dworkin, G. (1972). Paternalism. *The Monist, 56*(1), pp.64-84.

Eccles, J.C. (1972). *The Understanding of the Brain*. London: McGraw-Hill Book Company.

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press, 2016.

Epicurus, B. (2004). *Letter to Menoeceus*. University of Adelaide Library.

Feinberg, J. and Narveson, J. (1970). The Nature and Value of Rights. *The Journal of Value Inquiry, 4*(4), pp.243-260.

Felsen, G. and Reiner, P.B. (2011). How the Neuroscience of Decision Making Informs our Conception of Autonomy. *AJOB Neuroscience, 2*(3), pp.3-14.

Ferrari, P.F. and Coudé, G. (2018). Mirror Neurons, Embodied Emotions, and Empathy. In: Meyza, K.Z. ed., *Neuronal Correlates of Empathy: From Rodent to Human*. London: Academic Press.

Fischer, J.M. (2007). Compatibilism. In: Fischer, J.M., Kane, R., Pereboom, D. and Vargas, M. ed., *Four Views on Free Will*. Oxford: John Wiley & Sons.

Fischer, J.M. (2012). *Deep Control: Essays on Free Will and Value*. New York: Oxford University Press.

Fischer, J.M., Kane, R., Pereboom, D. and Vargas, M. (2007). *Four Views on Free Will*. Oxford: John Wiley & Sons.

Foucault, M. (1972). Truth and Power. In: Faubion, J.D. ed., *Power: The Essential Works of Foucault 1954 – 1984*. London: Penguin Books, 1994.

Frankfurt, H.G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy, 66*(23), pp.829-839.

Frankfurt, H.G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy, 68*(1), pp.5-20.

Freeman, S. (2003). *The Cambridge Companion to Rawls*. Cambridge: Cambridge University Press.

Freud, S. (1920). *Beyond the Pleasure Principle*. In: Reddick, J., ed., 2003. London: Penguin Books.

Freud, S. (1962). *The Ego and the I.D.*. New York: Norton.

Friederici, A.D., Makuuchi, M. and Bahlmann, J. (2009). The Role of the Posterior Superior Temporal Cortex in Sentence Comprehension. *Neuroreport, 20*(6), pp.563-568.

Galston, W.A. (1995). Two Concepts of Liberalism. *Ethics, 105*(3), pp.516-534.

Gazzaniga, M. S. (2005). *The Ethical Brain: The Science of our Moral Dilemmas*. Washington, D.C.: Dana Press

Gibbs, J.C. (2013). *Moral Development and Reality: Beyond the Theories of Kohlberg, Hoffman, and Haidt*. Oxford University Press.

Gibbs, J.C., Basinger, K.S., Grime, R.L. and Snarey, J.R. (2007). Moral Judgment Development Across Cultures: Revisiting Kohlberg's Universality Claims. *Developmental Review, 27*(4), pp.443-500.

Glannon, W. (2007). *Bioethics and the Brain*. Oxford: Oxford University Press.

Gomes, G. (1999). Volition and the Readiness Potential. *Journal of Consciousness Studies, 6*(8-9), pp.59-76.

Graybiel, A.M. (2000). The Basal Ganglia. *Current Biology, 10*(14), pp.R509-R511.

Gutmann, A. (1980). Children, Paternalism, and Education: A Liberal Argument. *Philosophy & Public Affairs, 9*(4), pp.338-358.

Habermas, J. (1990). Justice and Solidarity: On the Discussion Concerning Stage 6. In: Wren, T.E., ed., *The Moral Domain: Essays in the Ongoing Discussion Between Philosophy and the Social Sciences*. Cambridge: The MIT Press.

Haggard, P. (2008). Human Volition: Towards a Neuroscience of Will. *Nature Reviews Neuroscience, 9*(12), pp.934-946.

Haggard, P., Clark, S. and Kalogeras, J. (2002). Voluntary Action and Conscious Awareness. *Nature Neuroscience, 5*(4), pp.382-385.

Haggard, P. and Magno, E. (1999). Localising Awareness of Action with Transcranial Magnetic Stimulation. *Experimental Brain Research, 127*(1), pp.102-107.

Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science, 316*(5827), pp.998-1002.

Haidt, J. (2008). Morality. *Perspectives on Psychological Science, 3*(1), pp.65-72.

Hampton, J. (1989). Should Political Philosophy be done without Metaphysics? *Ethics, 99*(4), pp.791-814.

Hansard Society (2019). *Audit of Political Engagement 16: The 2019 Report*. London. PDF available at: assets.ctfassets.net/rdwvqctnt75b/7iQEHtrkIbLcrUkduGmo9b/cb429a657e97cad61e61853c05c8c4d1/Hansard-Society__Audit-of-Political-Engagement-16__2019-report.pdf (Accessed 01.10.2020).

Harkness, S., Edwards, C.P. and Super, C.M. (1981). *Social Roles and Moral Reasoning: A Case Study in a Rural African Community. Developmental Psychology, 17*(5), pp.595-603.

Harris, S. (2012). *Free Will*. New York: Free Press.

Hegel, G.W.F. (1821). Philosophy of Right. In: Knox, T.M., ed., *Hegel's Philosophy of Right*. Oxford: Clarendon Press, 1942.

Hobbes, T. (1651). *Leviathan*. Macpherson, C.B., ed.. London: Penguin, 1985.

Hobbes, T. Of Liberty and Necessity. In: Chappell, V., ed., *Hobbes and Bramhall on Liberty and Necessity*. Cambridge: Cambridge University Press, 1999.

Honderich, T. (1973). One Determinism. In: Honderich, T. ed., *Essays on Freedom of Action*. London: Routledge & Kegan Paul Ltd.

Honderich, T. (2002). *How Free Are You? The Determinism Problem*. Oxford: Oxford University Press.

Hume, D. (1738). *A Treatise on Human Nature*. Norton, D.F. & Norton, M.J. ed.. Oxford: Clarendon Press, 2007.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Millican, P., ed.. Oxford: Oxford University Press, 2008.

Hurley, T. (2003). John Rawls and Liberal Neutrality. In: Kukathas, C. Ed., John Rawls: *Critical Assessments of Leading Political Philosophers: Volume 4 – Political Liberalism and The Law of Peoples*. London: Routledge.

Johnson, O.A. (1977). Autonomy in Kant and Rawls: A Reply. *Ethics, 87*(3), pp.251-254.

Jønch-Clausen, K. and Kappel, K. (2016). Scientific Facts and Methods in Public Reason. *Res Publica, 22*(2), pp.117-133.

Jordan, M.D. (2016). *Teaching Bodies: Moral Formation in the Summa of Thomas Aquinas*. Oxford University Press.

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

Kane, R. (2000). The Dual Regress of Free Will and the Role of Alternative Possibilities. *Philosophical Perspectives, 14*, pp.57-79.

Kane, R. (2007). Libertarianism. In: Fischer, J.M., Kane, R., Pereboom, D. and Vargas, M. ed., *Four Views on Free Will*. Oxford: John Wiley & Sons.

Kane, R. (2011). Rethinking Free Will: New Perspectives on an Ancient Problem. In: Kane, R. ed., *The Oxford Handbook of Free Will: Second Edition*. Oxford: Oxford University Press.

Kant, I. (1787). *Critique of Pure Reason*. Translated by Marcus Weigelt. London: Penguin Classics, 2007.

Kant, I. (1788). *Critique of Practical Reason*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 2015.

Kant, I. (1797). *The Moral Law: Groundwork of the Metaphysic of Morals*. Translated by H.J. Paton. Abingdon: Routledge, 2005.

Kazén, M., Baumann, N. and Kuhl, J. (2003). Self-Infiltration vs. Self-Compatibility Checking in Dealing with Unattractive Tasks: The Moderating Influence of State vs. Action Orientation. *Motivation and Emotion, 27*(3), pp.157-197.

Keller, I. and Heckhausen, H. (1990). Readiness Potentials Preceding Spontaneous Motor Acts: Voluntary vs. Involuntary Control. *Electroencephalography and Clinical Neurophysiology, 76*(4), pp.351-361.

Kingsley, R.E. (2000). *Concise Text of Neuroscience*. London: Lippincott Williams & Wilkins.

Knight, C. (2017). Reflective Equilibrium. In: Blau, A., ed., *Methods in Analytical Political Theory*, Cambridge: Cambridge University Press.

Kohlberg, L. and Hersh, R.H. (1977). Moral Development: A Review of the Theory. *Theory Into Practice, 16*(2), pp.53-59.

Kornhuber, H.H. & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente

Potentiale. *Pflüger's Archiv für die Gesamte Physiologie des Menschen und der Tiere, 284*(1), pp.1-17.

Kuhl, J. & Koole, S.L. (2004). The Workings of the Will: A Functional Approach. In: Greenberg, J., Koole, S.L., Psyzczynski, T., ed., *Handbook of Experimental Existential Psychology*. London: The Guilford Press.

Kukathas, C. (2003). General Introduction. In: Kukathas, C. Ed., *John Rawls: Critical Assessments of Leading Political Philosophers: Volume 1 – Foundations and Methods*. London: Routledge.

Laplace, P.S. (1814). A Philosophical Essay on Probabilities. In: Dale, A.I. ed., *Pierre-Simon Laplace: Philosophical Essay on Probabilities – Translated from the Fifth French Edition of 1825 With Notes by the Translator*. New York: Springer-Verlag, 2012.

Larmore, C. (1996). *The Morals of Modernity*. Cambridge University Press.

Leisman, G., Macahdo, C., Melillo, R. and Mualem, R. (2012). Intentionality and "Free-Will" from a Neurodevelopmental Perspective. *Frontiers in Integrative Neuroscience, 6*, pp.1-12.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), pp.529-539.

Libet, B. (1999). Do We Have Free Will? In: Libet, B., Freeman, A., & Sutherland, K., ed. *The Volitional Brain: Towards a Neuroscience of Free Will*. Exeter: Imprint Academic.

Libet, B. (2006). Reflections on the Interaction of the Mind and Brain. *Progress in Neurobiology*, *78*(3-5), pp.322-326.

Libet, B., Gleason, C.A., Wright, E.W. and Pearl, D.K. (1983). Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act. *Brain, 106*(3), pp.623-642.

Locke, J. (1690). *An Essay Concerning Human Understanding*. Woolhouse, R., ed.. London: Penguin, 1998.

Longstaff, A. (2005). *Neuroscience: Second Edition*. Abingdon: Taylor & Francis Group.

Luther, M. (1525). *The Bondage of the Will*. Translated by Packer, J.I. & Johnston, O.R. Grand Rapids: Baker Academic, 1957.

Macpherson, C.B. (1985). Introduction. In: Hobbes, T. *Leviathan*. London: Penguin Books.

Mandle, J. (2009). *A Theory of Justice: An Introduction*. Cambridge: Cambridge University Press.

Marx, K. (1843). On the Jewish Question. In: Easton, L. D., Gudat, K.H., eds., *Writings of the Young Marx on Philosophy and Society*. New York: Doubleday.

Masmoudi, R.A. (2003). What is liberal Islam? The Silenced Majority. *Journal of Democracy, 14*(2), pp.40-44.

Mason, M., & Gibbs, J. C. (1993a). Role-Taking Opportunities and the Transition to Advanced Moral Judgment. *Moral Education Forum, 18*(3) pp.1–12.

Mccall, S. (2013). Does the Brain Lead the Mind? *Philosophy and Phenomenological Research, 86*(2), pp.262-265.

McDonald, N.M. & Messinger, D.S. (In Press) (2011). In Acerbi, A., Lombo, J.A., & Sanguineti, J.J., ed., *Free will, Emotions, and Moral Actions: Philosophy and Neuroscience in Dialogue*. IF-Press. Available at: http://www.psy.miami.edu/faculty/dmessinger/c_c/rsrcs/rdgs/emot/McDonald-Messinger_Empathy%20Development.pdf (Accessed 26.10.2018).

Mele, A.R. (2006). *Free Will and Luck*. Oxford: Oxford University Press.

Mele, A. (2014). *Free: Why Science Hasn't Disproved Free Will*. Oxford: Oxford University Press.

Michelman, F.I. (1993). The Subject of Liberalism. *Stanford Law Review, 46*, p.1807-1833.

Mill, J. (1859). On Liberty. In: Gray, J. ed., *On Liberty and other Essays*. Oxford: Oxford University Press, 1991.

Miller, R. (1974). Rawls and Marxism. *Philosophy & Public Affairs, 3*(2), pp.167-191.

Mink, J.W. (2003). The Basal Ganglia and Involuntary Movements: Impaired Inhibition of Competing Motor Patterns. *Archives of neurology, 60*(10), pp.1365-1368.

Moore, G.E. (1947). *Ethics*. Oxford: Oxford University Press.

Morse, S.J. (2008). Determinism and the Death of Folk Psychology: Two Challenges to Responsibility from Neuroscience. *Minn. JL Sci. & Tech., 9*, pp.1-36.

Mumford, S. & Tugby, M. (2013). Introduction: What is the Metaphysics of Science? In: Mumford, S. & Tugby, M., ed., *Metaphysics and Science*. Oxford: Oxford University Publishing.

Murphy, A.R. (1998). Rawls and a Shrinking Liberty of Conscience. *The Review of Politics*, *60*(2), pp.247-276.

Musumeci, G., Castorina, S., Castrogiovanni, P., Loreto, C., Leonardi, R., Aiello, F.C., Magro, G. and Imbesi, R. (2015). A Journey Through the Pituitary Gland: Development, Structure and Function, with Emphasis on Embryo-Foetal and Later Development. *Acta Histochemica, 117*(4-5), pp.355-366.

Neshige, R., Lüders, H. and Shibasaki, H. (1988). Recording of Movement-Related Potentials From Scalp and Cortex in Man. *Brain: A Journal of Neurology*, 111, pp.719-736.

Noddings, N. (1994). Conversation as Moral Education. *Journal of Moral Education, 23*(2), pp.107-118.

Nozick, R. (1974). *Anarchy, State, Utopia*. Oxford: Blackwell.

Nucci, L, Narvaez, D., and Krettenauer, T. (2014). Introduction and Overview. In: Nucci, L, Narvaez, D., and Krettenauer, T., ed., *Handbook of Moral and Character Education*. Abingdon: Routledge.

Nussbaum, M.C. (2011a). Rawls's Political Liberalism. A Reassessment. *Ratio Juris, 24*(1), pp.1-24.

Nussbaum, M.C. (2011b). Perfectionist Liberalism and Political Liberalism. *Philosophy & Public Affairs, 39*(1), pp.3-45.

Olsaretti, S. (1998). Freedom, Force and Choice: Against the Rights-Based Definition of Voluntariness. *Journal of Political Philosophy, 6*(1), pp.53-78.

Parfit, D. (2011). *On What Matters: Volume One (Vol. 1)*. Oxford: Oxford University Press.

Parisette-Sparks, A., Bufferd, S.J. and Klein, D.N. (2017). Parental Predictors of Children’s Shame and Guilt at Age 6 in a Multimethod, Longitudinal Study. *Journal of Clinical Child & Adolescent Psychology, 46*(5), pp.721-731.

Pereboom, D. (2006). *Living without Free Will*. Cambridge University Press.

Pereboom, D. (2007). Hard Incompatibilism. In: Fischer, J.M., Kane, R., Pereboom, D. and Vargas, M. ed., *Four Views on Free Will*. Oxford: John Wiley & Sons.

Pereboom, D. (2011). Free-Will Skepticism and Meaning in Life. In: Kane, R., ed., *The Oxford Handbook of Free Will: Second Edition*. New York: Oxford University Press.

Pettit, P. (2007). Neuroscience and Agent-Control. In: Ross, D., Spurrett, D., Kinkaid, H., & Lynn Stephens, G., ed., *Distributed Cognition and the Will*. Cambridge: The MIT Press.

Pöppel, E. (1978). Time perception. In: Held, R., Leibowitz, H. W., Teuber, H.L., ed., *Perception*. Berlin: Springer.

Popper, K. (1963). The Demarcation Between Science and Metaphysics. In: Popper, K., ed., *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge and Kegan Paul Limited.

Porges, S.W. and Lewis, G.F. (2010). The Polyvagal Hypothesis: Common Mechanisms Mediating Autonomic Regulation, Vocalizations and Listening. In: *Handbook of Behavioral Neuroscience, Vol. 19*. Elsevier.

Postle, B.R. (2015). *Essentials of Cognitive Neuroscience*. Chichester: John Wiley & Sons, Ltd.

Praamstra, P., Stegeman, D.F., Horstink, M.W.I.M. and Cools, A.R. (1996). Dipole Source Analysis Suggests Selective Modulation of the Supplementary Motor Area Contribution to the Readiness Potential. *Electroencephalography and Clinical Neurophysiology, 98*(6), pp.468-477.

Puka, B. (1990). The Majesty and Mystery of Kohlberg's Stage 6. In: Wren, T.E., ed., *The Moral Domain: Essays in the Ongoing Discussion Between Philosophy and the Social Sciences*. Cambridge: The MIT Press.

Quong, J. (2011). *Liberalism Without Perfection*. Oxford: Oxford University Press.

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Rawls, J. (1974). The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association, 48*(1975), pp. 5-22.

Rawls, J. (1980). Rational and Full Autonomy. *The Journal of Philosophy, 77*(9), pp.515-535.

Rawls, J. (1982a). The Basic Liberties and their Priority. *The Tanner Lectures on Human Values, 3*, pp.3-87.

Rawls, J. (1982b). Social Unity and Primary Goods. In: Sen, A. & Williams, B., ed., *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.

Rawls, J. (1985). Justice as Fairness: Political not Metaphysical. *Philosophy & Public Affairs*, *14*(3) pp.223-251.

Rawls, J. (1988). The Priority of Right and Ideas of the Good. *Philosophy & Public Affairs*, pp.251-276.

Rawls, J. (1989). The Domain of the Political and Overlapping Consensus. *NYUL Rev., 64*, pp.233-255.

Rawls, J. (1993). The Law of Peoples. *Critical Inquiry, 20*(1), pp.36-68.

Rawls, J. (2000). *Lectures on the History of Moral Philosophy*. Cambridge: Harvard University Press.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Cambridge: Harvard University Press.

Rawls, J. (2005). *Political Liberalism: Expanded Edition*. New York: Columbia University Press.

Rawls, J. (2007). *Lectures on the History of Political Philosophy*. Cambridge: Harvard University Press.

Raz, J. (1986). *The Morality of Freedom*. Oxford: Oxford University Press.

Raz, J. (1990). Facing diversity: The Case of Epistemic Abstinence. *Philosophy & Public Affairs, 19*(1), pp.3-46.

Reader, S. (2007). The Other Side of Agency. *Philosophy, 82*(4), pp.579-604.

Rist, J.M. (1969). Augustine on Free will and Predestination. *The Journal of Theological Studies, 20*(2) pp.420-447.

Rolls, E.T. (2015). Limbic Systems for Emotion and for Memory, but No Single Limbic System. *Cortex, 62*, pp.119-157.

Ross, D., Ladyman, J. and Spurrett, D. (2007). In Defence of Scientism. In: Ladyman, J., Ross, D., ed., *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Roth, G. (2003a). Between Cortex and the Basal Ganglia: Neuroscientific Accounts of Voluntary Action. In: Maasen, S., Prinz, W., & Roth, G., ed., *Voluntary Action: Brain, Minds, and Sociality*. Oxford: Oxford University Press.

Roth, G. (2003b). The Interaction of Cortex and Basal Ganglia in the Control of Voluntary Actions. In: Maasen, S., Prinz, W., & Roth, G., ed., *Voluntary Action: Brain, Minds, and Sociality*. Oxford: Oxford University Press.

Roth, G. (2010). Free Will: Insights from Neurobiology. In: Frey, U.J., Störmer, C., Willführ, K.P., ed., Homo Novus: A Human Without Illusions. London: Springer.

Rowe, W.L. (1964). Augustine on foreknowledge and free will. *The Review of Metaphysics, 18(2)*, pp.356-363.

Rousseau, J.J. (1762). *The Social Contract*. Translated by Peter Constantine. London: Penguin, 1998.

Ruckstaetter, J., Sells, J., Newmeyer, M.D. and Zink, D. (2017). Parental Apologies, Empathy, Shame, Guilt, and Attachment: a Path Analysis. *Journal of Counseling & Development, 95*(4), pp.389-400.

Ruhnau, E. (1995). Time-Gestalt and the Observer. In: Metzinger, T., ed., *Conscious Experience*. Thorverton: Imprint Academic.

Russell, B. (1946). *History of Western Philosophy*. (Published 2004). Abingdon: Routledge.

Salat, D.H., Buckner, R.L., Snyder, A.Z., Greve, D.N., Desikan, R.S., Busa, E., Morris, J.C., Dale, A.M. and Fischl, B. (2004). Thinning of the Cerebral Cortex in Aging. *Cerebral Cortex*, 14(7), pp.721-730.

Sandel, M.J. (1984). The Procedural Republic and the Unencumbered Self. *Political Theory*, 12(1), pp.81-96.

Sasaki, K. & Gemba, H. (1991). Cortical Potentials Associated with Voluntary Movements in Monkey. In: Brunia, C.H.M., Mulder, G. & Verbaten, M.N., ed., *Event-Related Brain Research (EEG Suppl.42)*. Amsterdam: Elsevier.

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

Scheffler, S. (1994). The Appeal of Political Liberalism. *Ethics, 105*(1), pp.4-22.

Schopenhauer, A. (1839). *On the Freedom of the Will*. Translated by Konstantin Kolenda, 1985. Second Edition. Oxford: Basil Blackwell, Ltd.

Scott Johnston, J. (2005). Rawls's Kantian Educational Theory. *Educational Theory, 55*(2), pp.201-218.

Searle, J.R. (2000). Consciousness, Free Action and the Brain. *Journal of Consciousness Studies, 7*(10), pp.3-22.

Sen, A.K. (1999). *Development as Freedom*. Oxford: Oxford University Press.

Sen, A.K. (2009). *The Idea of Justice*. Cambridge: Harvard University Press.

Shapiro, D. (1995). Why Rawlsian Liberals Should Support Free Market Capitalism. *Journal of Political Philosophy, 3*(1), pp.58-85.

Shibasaki, H. & Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology, 117*(11), pp.2341-2356.

Sidgwick, H. (1888). The Kantian Conception of Free Will. *Mind, 13*(51), pp.405-412.

Simkulet, W. (2013). Autonomy as Free Will. *AJOB Neuroscience, 4*(4), pp.71-72.

Skinner, B.F. (1971). *Beyond Freedom and Dignity*. Indianapolis: Hackett Publishing, 2002.

Smart, J.J.C. (1961). Free-will, Praise and Blame. *Mind, 70*(279), pp.291-306.

Smilansky, S. (1994). The Ethical Advantages of Hard Determinism. *Philosophy and Phenomenological Research, 54*(2), pp.355-363.

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

Smilansky, S. (2003). Free will, Egalitarianism and Rawls. *Philosophia, 31*(1-2), pp.127-138.

Snarey, J. R. (1985). The Cross-Cultural Universality of Social-Moral Development: A Critical Review of Kohlbergian Research. *Psychological Bulletin, 97*, 202–232.

Soon, C.S., Brass, M., Heinze, H.J. and Haynes, J.D. (2008). Unconscious Determinants of Free Decisions in the Human Brain. *Nature Neuroscience, 11*(5), p.543.

Spinoza, B.D. (1677). *Ethics*. Translated by Edwin Curley. London: Penguin Books, 1996.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Strawson, G. (2018). *Things That Bother Me: Death, Freedom, The Self, etc.*. New York: New York Review of Books.

Strike, K.A. (1994). On the Construction of Public Speech: Pluralism and Public Reason. *Educational Theory, 44*(1), pp.1-26.

Stump, E. (1996). Libertarian Freedom and the Principle of Alternative Possibilities. In: Jordan, J. & Howard-Snyder, D., ed., *Faith, Freedom, and Rationality: Philosophy of Religion Today*. London: Rowman & Littlefield.

Stump, E. (1997). Aquinas's Account of Freedom: Intellect and Will. *The Monist, 80*(4), pp.576-597.

Suhler, C.L. and Churchland, P. (2011). Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience, 23*(9), pp.2103-2116.

Thompson, R.F. (2000). *The Brain: A Neuroscience Primer, Third Edition*. New York: Worth Publishers.

Titelbaum, M.G. (2008). What would a Rawlsian Ethos of Justice look like? *Philosophy & Public Affairs, 36*(3), pp.289-322.

Van Inwagen, P. (2018). *Metaphysics*. Abingdon: Routledge.

Van Inwagen, P. and Griffiths, A.P. (1985). *An Essay on Free Will*. Oxford: Clarendon Press.

Vargas, M. (2004). Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly, 85*(2), pp.218-241.

Velmans, M. (1991). Is Human Information Processing Conscious? *Behavioral and Brain Sciences, 14*(4), pp.651-669.

Waldron, J. (2002). *God, Locke, and Equality: Christian Foundations in Locke's Political Thought*. Cambridge University Press.

Ward, J. (2015). *The Student's Guide to Cognitive Neuroscience: Third Edition*. Hove: Psychology Press.

Watt, W.M. (1946). Free will and Predestination in Early Islam. *The Muslim World, 36*(2), pp.124-152.

Weatherford, R. (1991). *The Implications of Determinism*. London: Routledge.

Wegner, D. (2002). *The Illusion of Conscious Will*. London: The MIT Press.

Whitty, P.F., Owoeye, O. and Waddington, J.L. (2008). Neurological Signs and Involuntary Movements in Schizophrenia: Intrinsic to and Informative on Systems Pathobiology. *Schizophrenia Bulletin, 35*(2), pp.415-424.

Widerker, D. (2000). Frankfurt's Attack on the Principle of Alternative Possibilities: A Further Look. *Philosophical Perspectives, 14*, pp.181-201.

Wike, R. & Schumacher, S. (2020). Democratic Rights Popular Globally but Commitment to Them Not Always Strong. *Pew Research Center*. Washington, D.C.. (Available at: www.pewresearch.org/global/2020/02/27/democratic-rights-popular-globally-but-commitment-to-them-not-always-strong/)

Williams, B. (1985). *Ethics and the Limits of Philosophy*. London: Fonatana Press.

Wilson, E.O. (2000). *Sociobiology*. Harvard University Press.

Wilson, R., Gaines, J. and Hill, R.P. (2008). Neuromarketing and Consumer Free Will. *Journal of Consumer Affairs, 42*(3), pp.389-410.

Yazawa, S., Ikeda, A., Kunieda, T., Ohara, S., Mima, T., Nagamine, T., Taki, W., Kimura, J., Hori, T. and Shibasaki, H. (2000). Human Presupplementary Motor Area is Active Before Voluntary Movement: Subdural Recording of Bereitschaftspotential from Medial Frontal Cortex. *Experimental Brain Research, 131*(2), pp.165-177.

Yoder, K.J. and Decety, J. (2018). The Neuroscience of Morality and Social Decision-Making. *Psychology, Crime & Law, 24*(3), pp.279-295.

Young, I.M. (1995). Rawls's Political Liberalism. *Journal of Political Philosophy, 3*(2), pp.181-190.

Zimmerman, K.A. (2017). What is Short-Term Memory Loss? *Live Science*. Available at: https://www.livescience.com/42891-short-term-memory-loss.html (Accessed 26.10.2018).