Rational Reflection and An Integrative Account of Moral Cognition

A thesis submitted to The University of Otago in fulfilment of the requirements for the Bachelor of Medical Science (Hons) degree

February 25, 2019

Finn Whittington

University of Otago

Abstract

This thesis is about the relationship between the brain and morality. In the last two decades there has been a rapid growth in the neuroscience and psychology of morality, such that a new a field has emerged called "moral cognition". However, much of this study is reductive in nature and philosophically uninformed. The aim of this thesis is to discuss the need for an account of moral cognition that is integrative in its character and interdisciplinary in its approach. The current neuroscientific literature shows that moral cognition operates as a physically integrated network. No single part of the brain is solely responsible for moral processing. This raises the question of how the various regions and processes found to be associated with moral cognition function together. In particular, do some processes such as emotion or reason take precedence over the other? To answer these questions a neuroscience informed by moral philosophy is required. Indeed, the study of moral cognition is necessarily dependent on moral philosophical theories, as a notion of what counts as "morality" is required prior to any study, in order to give the study a clear focus. Some of the key figures in the current study of moral cognition are aware of this, and often explicitly take up philosophical presuppositions. However, they do so in cursory ways. Furthermore, current neuroscientific techniques are unable to describe the temporal and holistic nature of moral cognition. Any analysis which relies solely on these techniques is missing normal aspects of moral thinking. Therefore, neuroscience can only be part of an interdisciplinary approach. A theoretical neuroscience informed by a nuanced moral philosophy is needed to start to approach a description of the relationship between the brain and morality. This relationship between neuroscience and moral science can be characterised as involving two perspectives. The scientific perspective brings with it an explanatory account, and is concerned with causal descriptions of brain processes underlying moral behaviour. At the same time, a nuanced moral philosophy inevitably identifies the significance of a first person or agent perspective. From this perspective morality involves a sense of obligation – it recognises that morality is normative. Neuroscience cannot describe morality from a first-person or normative perspective, but the normative aspects of morality cannot be discounted. In this thesis the moral philosophy of Christine Korsgaard is discussed as it is particularly effective in introducing the first-person or normative account of morality and has the advantage of a neuroscientifically plausible view of reason and reflection. The thesis will finish with

an exercise of interdisciplinary analysis, where neuroscience can reciprocally inform moral philosophy. I will discuss key aspects of Korsgaard's theory in the light of contemporary themes in current neuroscience. Specifically, I discuss how her understanding of rational reflection is both supported and moderated by the work of contemporary neuroscience, particularly the work of Antonio Damasio. I also show how her theory and neuroscience can jointly explain the integration of emotion and reason in moral judgement.

Acknowledgements

I had known for a long time that I needed to take time out of medicine to focus on philosophy, neuroscience and bioethics. Writing this thesis has been challenging and gruelling at times, but most significantly it has been enlightening. I have a great many people to thank for supporting me through the difficult times and exciting and enlightening me about the world and philosophy. I owe a tremendous amount of gratitude to Simon Walker, my primary supervisor, a dedicated mentor and teacher, who would always give time to listen and discuss, and who helped me to (finally) understand Kant. I also owe my deepest gratitude to my two secondary supervisors, Neil Pickering and Liz Franz. Thank you, Neil, for teaching me how to build a strong argument and for always having poignant examples or anecdotes at the ready. Thank you, Liz, for sharing your insight and knowledge about neurons, the brain and all the circuits in between. I also owe my sincere gratitude to Grant Gillett, who showed me the pathway between medicine and neurophilosophy and kicked off my research journey. To the students and staff at the bioethics centre, for all your support, input, care and conversations. To Jacqui Bradshaw at the Med school for your direction and being so friendly. To Angela and Corey who made the bioethics centre so open. To Phyllis Paykel and the organisers of her estate, for providing me with this opportunity and financial support in the form of a scholarship. To Hunter, for the unconditional love. To Dad, for worrying for me and my brother, the unending support and all the care packages. To Amy, for your love and kindness and the kindness you foster in me and in others. To Seth, for always being up for a deep philosophical chat, and sharing all the music, art, films and politics. To Danushi, for the love and the journey we have been through together.

Table of Contents

Table of Figures

Introduction

Kant's second critique, the *Critique of Practical* Reason, is a systematic account of morality and its central role in human life. This passage from its closing pages describes a kind of "science" aimed at developing moral "wisdom":

> …We have at hand examples of the morally judging reason. We may analyse them into their elementary concepts, adopting, in default of mathematics, a process similar to that of chemistry, i.e., we may, in repeated experiments on common sense, separate the empirical from the rational, exhibit each of them in a pure state, and show what each by itself can accomplish. Thus we shall avoid the error of a crude and unpracticed judgment and (something far more important) the extravagances of genius, by which, as by the adepts of the philosopher's stone, visionary treasures are promised and real treasures are squandered for lack of methodical study and knowledge of nature. In a word, science (critically sought and methodically directed) is the narrow gate that leads to the doctrine of wisdom, when by this is understood not merely what one ought to do but what should serve as a guide to teachers in laying out plainly and well the path to wisdom which everyone should follow, and in keeping others from going astray. It is a science of which philosophy must always remain the guardian; and though the public takes no interest in its subtle investigations, it may very well take an interest in the doctrines which such considerations first make clear to it.
>
> (Kant, 1993, pp. 170-171)

Kant is looking ahead to what he sees as the future of ethics. Living at the time of the enlightenment, he recognised the extensive and rapid development of the physical sciences, which had only recently shrugged off the shroud of "superstition". For instance, a century before Kant's writing, Newton had made great strides by describing gravity and the laws of motion. Such scientific advances uncovering "the starry heavens above" filled Kant with awe (Kant, 1993, p. 169). What was next, in his mind, was to investigate the equally wonderous "moral law within". Vital was the further excision of superstition. In this passage, Kant proposes that morality requires a systematic and methodical study, not unlike science, where the "empirical" and the "rational" would be used effectively and synergistically.

The scientific frontiers in Kant's time were mechanical physics, chemistry, astronomy and anatomy. In modern times, with the development of cognitive neuroscience, the human mind, and with it human morality, has begun to be investigated as a physical science. Within neuroscience, a field has emerged which specifically studies the relationship between the brain and moral thinking. It appears that Kant's interest in the starry heavens, a stand-in for scientific mystery, has begun to collide with the study of the "moral law".

This new scientific study of moral cognition includes the psychological and brain processes involved in human moral thinking. This includes the neural processes underlying moral judgements, moral emotions and moral motivations. Individual researchers and philosophers each employ specific definitions of what these are, that segregate morally relevant behaviours from other behaviours in terms of their effects (like increasing co-operation) or motivations (necessity of altruism). This diversity in conceptualising moral cognition presents a significant challenge for the neuroscientific study of moral cognition. However, in all cases the broad aim of studying moral cognition is to explain the relationship between the brain and the operation of "morality".

In the past two decades, understandings of moral cognition have undergone significant development, largely due to development in neuroimaging techniques in cognitive neuroscience, such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and transcranial magnetic stimulation (TMS). Such technologies have enabled experimental work to provide extensive data about the brain regions and processes involved in various kinds of moral thinking and behaviour. Models of moral cognition have developed alongside this experimental work, which attempt to systematically describe principles and processes common to moral thinking. Often included are explanations of the operative roles of emotion, reason and social cognition in moral judgements and motivations. These modern models often have roots in older theories from moral philosophy, such as Kantian ethics, utilitarianism, virtue theory and sentimentalism.

In this thesis I discuss the need for an account of moral cognition that is integrative in character and interdisciplinary in approach. I will argue that we need such

a model to understand the relationship between the brain and our moral lives. The model proposed is "integrative" in three ways:

1. Neuroscientifically integrated: Moral cognition refers to a network of brain regions. These regions operate with integrative brain processes and as such should be described using theories of higher-level cognition. The intertwinement of reason and emotion is essential in this framework.
2. Integrated across disciplines (as in "inter-disciplinary"): From a methodological standpoint any comprehensive description of moral cognition needs to refer to moral philosophy, neuroscience and psychology (at least) and integrate their respective methods, generating a synthesis of perspectives.
3. Integrated explanatory and normative accounts of morality: To form a complete account of our moral lives, both explanatory and normative descriptions are necessary.

By an explanatory account of morality, I am referring to a causal description of morality, typical to psychology and neuroscience. By a normative account of morality, I am referring to an explanation of the normativity underlying moral claims, i.e. a justification of the "ought" inherent to such claims.

If one wants to describe moral cognition and the relationship between the brain and our moral lives accurately and comprehensively, a framework based on these three features is required. I will argue for this requirement in two ways. Firstly, I will critically analyse the neuroscience of moral cognition using moral philosophy. Secondly, I will undertake an interdisciplinary analysis, including a reciprocal critique of moral philosophy using neuroscience.

The current neuroscientific literature shows that moral cognition involves a coordination of brain processes, which operate as a physically integrated network. This is sufficient to show that there is no part of the brain solely responsible for moral processing. Rather moral cognition is distributed and multi-faceted, involving emotion, reason and social cognition. This raises the question of how the various regions and processes found to be associated with moral cognition function together. In particular, do some processes such as emotion or reason take precedence over the other? To decide which is the best account of moral cognition, requires a neuroscientist to turn outside the world of neuroscience, and to the world of moral philosophy.

The study of moral cognition is necessarily dependent on moral philosophical theories, as a notion of what counts as "morality" and "moral judgement" are required prior to any study, in order to give the study a clear focus. Some of the key figures in the current study of moral cognition are aware of this, and often explicitly take up philosophical presuppositions. However, they do so in cursory ways. Furthermore, current neuroscientific techniques are unable to describe the temporal and holistic nature of moral cognition. Any analysis which relies solely on these techniques is missing normal aspects of moral thinking. Therefore, neuroscience can only be part of an interdisciplinary approach. A theoretical neuroscience informed by a nuanced moral philosophy is needed to start to approach a description of the relationship between the brain and morality.

This relationship between neuroscience and moral science can be characterised as involving two perspectives. The scientific perspective brings with it an explanatory account and is concerned with causal descriptions of brain processes underlying moral behaviour. At the same time, a nuanced moral philosophy inevitably identifies the significance of a first person or agent perspective. From this perspective morality involves a sense of obligation – it recognises that morality is normative. Neuroscience cannot describe morality from a first-person or normative perspective, but the normative aspects of morality cannot be discounted. In this thesis the moral philosophy of Christine Korsgaard is discussed as it is particularly effective in introducing the first-person or normative account of morality and has the advantage of giving a neuroscientifically plausible view of reason and reflection.

Neuroscience, if it is to further a study of moral cognition needs to recognise the shortcomings and limits of the presuppositions made about the nature of morality, some of which come out in experimental restraints and others through a cursory use of moral philosophy. It is worthwhile for the neuroscience to be developed in this way, so that it performs its descriptive role as best as possible. If we do this well it will also set reasonable limits to any normative account of morality from the perspective of moral philosophy.

This thesis will finish with an exercise of interdisciplinary analysis, where neuroscience will reciprocally inform moral philosophy. I discuss key aspects of Korsgaard's theory in the light of contemporary themes in current neuroscience.

Specifically, I discuss how her understanding of rational reflection is both supported and moderated by the work of contemporary neuroscience, particularly the work of the neurologist Antonio Damasio. From this I will develop an integrative model of moral reasoning, which will show how both emotion and reasons are integrated in our moral judgement. I will briefly indicate how this relates to certain ongoing debates in neuroscience. Importantly this framework is not intended to be a comprehensive or exclusive model of moral cognition, rather a basic framework for an integrative account. As such, certain philosophical perspectives will remain unaddressed and many questions will remain open.

In a review by Wagner et al. (2017) the field of moral cognition was analysed from the perspective of bioethics, with a particular focus on the normative implications of neuroscientific work. They found a poverty in theoretical work examining such implications of moral cognition, and strongly recommended more work in this field from the perspective of bioethics and moral philosophy. This thesis is in line with this recommendation.

Theoretical work in this field is incredibly important to further develop an understanding of moral cognition, which can aid in the advancement of cognitive neuroscience more generally. Because of the complex nature of moral cognition, where the component brain regions function as a network, the neuroscientist Joshua Greene (2015) argues that it represents the perfect testing ground for theories of higher-level cognition. This is evidenced by a recent review by Gillett and Franz (2014) who applied John Hughlings-Jackson's ideas about complex brain organisation to the field of moral cognition. Therefore, theoretical work in moral cognition is well poised to further general debates in cognitive neuroscience.

The study of moral cognition has intense and extensive implications for ethics. In short, knowing why we are subject to our moral intuitions can aid our ethical thinking. It can show us blind spots in our moral reasoning and can inform strategies in improving our moral judgments. For example, recently a debate concerning the utility of empathy in morality has intensified. Some researchers have argued for the deemphasis of empathy in policy development, citing its intrinsic partiality as a problem for justice and any attempt to maximise outcomes, see for example Paul Bloom (2016). Others maintain that empathy is foundational to morality and that any

deemphasis would be detrimental for ethics (Slote, 2007). By improving our understanding of empathy and its role in moral cognition, we can be more informed about its normative benefits or drawbacks. An integrative account, like that put forward in this thesis would greatly benefit this debate. It would argue that empathy should be viewed as part of a network as opposed to being viewed independently, in an attempt to more accurately determine its normative implications.

A well-known example that illustrates how knowledge about moral cognition may benefit ethics is Peter Singer's famous "drowning child" thought experiment. This thought experiment is intended to test our intuitions about the extent of our moral obligations. It involves the following challenge which he put to his students:

> I ask them to imagine that their route to the university takes them past a shallow pond. One morning, I say to them, you notice a child has fallen in and appears to be drowning. To wade in and pull the child out would be easy but it will mean that you get your clothes wet and muddy, and by the time you go home and change you will have missed your first class.
>
> (Singer, 1997)

Unanimously the students agree that they have an obligation to save the child. Singer then probes further and asks would this obligation remain if the child was in the same situation but far away, in a different country, and we could similarly help with little cost to ourselves. Again, the students agree. However, Singer (1997) says: "we are all in that situation of the person passing the shallow pond". For little cost to ourselves we can help those less fortunate, like those in famine in South Sudan or Yemen, for example. Singer's thought experiment was designed to show how ostensibly hypothetical ethical dilemmas are often tangible and highly practical. Despite Singer's (and the student's) reasoning illustrating the extent of our obligations, most of his students would backtrack and attempt to show how the thought experiment does not match with reality. This led Singer (1997) to claim that: "There is, of course, for many students and for various reasons a gap between acknowledging what we ought to do, and doing it…". If we can study the underpinnings of our moral thinking, then perhaps we can understand such a disconnect. We could formulate strategies to overcome such discrepancies in our moral reasoning. Some ethicists even go so far as advocating for human "moral enhancement", which could involve the use of "drugs, implants and biological

(including genetic) interventions…" to improve people's moral-decision making (Persson and Savulescu, 2008). A complete understanding of the brain processes underlying our moral motivations, reasoning and intuitions would allow us to precisely identify the neural targets of any enhancement and alter them effectively. Hence, an understanding of moral cognition is essential for this project.

In this thesis I discuss the need for an account of moral cognition that is integrative in its character and interdisciplinary in its approach. It is structured as follows: The aim of chapter one is to show the support for a physically integrative model of moral cognition in the neuroscientific literature. To achieve this, I will review recent neuroscientific studies in the field of moral cognition, with a focus on the implicated brain regions and their functions. A set of important axioms emerge from this research that any subsequent model of moral cognition must abide. These include that moral cognition is a multifaceted process involving a network of brain regions, disparate in space and function (implicating emotion, reason, and social cognitive processes).

The aim of chapter two is to discuss the need for an interdisciplinary analysis in the study of moral cognition. To achieve this, I will discuss three groups of challenges facing the neuroscience of moral cognition. These challenges show that the neuroscientific study of moral cognition is necessarily dependent on theories in moral philosophy, however, much of this study is reductive in nature and philosophically uninformed. A theoretical neuroscience informed by a nuanced moral philosophy is needed to adequately describe the relationship between the brain and morality.

In chapter three I will lay out the moral philosophy of Christine Korsgaard, as part of this interdisciplinary analysis.[1] Korsgaard separates accounts of morality into explanatory and normative components. I will show how neuroscientific accounts of moral cognition alone are inadequate to explain the normativity behind moral judgements. This further indicates the need for an interdisciplinary analysis in describing moral cognition.

In chapter four I will carry out an exercise of interdisciplinary analysis by developing an integrative account of moral cognition. This account of moral cognition

[1] Korsgaard's theoretical positions will primarily be taken from Korsgaard, C. (1996) *The sources of normativity.* Cambridge University Press.

will be based on an integration of moral philosophy and neuroscience. In particular, it is an integration of Korsgaard's ideas about reason and reflection with current empirical and theoretical neuroscience of moral cognition, including the work of Damasio. By integrating the differing accounts of moral decision-making of Korsgaard and Damasio, an account of morality, that adequately deals with the roles of feelings, empathy, reflection and practical reason will emerge. This account will constitute the framework of an integrated moral cognition, that emphasises the role of reflection, while criticising any hard distinction between emotion and reason. Instead I will argue for a feeling guided and cognitively informed model of practical reasoning.

1 The integrative moral brain

The overarching objective of this thesis is to discuss the need for an account of moral cognition that is integrative in character and interdisciplinary in approach. To do this it is necessary to have an understanding of the state of the neuroscientific literature concerning moral cognition. This is the focus of the present chapter. I will describe the brain regions commonly implicated in making moral judgements and draw some basic conclusions about the nature of the "moral brain". This discussion will set out the neuroscientific landscape before launching into specific critiques of the neuroscience in chapter two.

Much of the literature of moral cognition is concerned with describing brain regions and processes involved in the making of moral judgements. There is a general consensus that there is no dedicated moral module of the brain (Gillett and Franz, 2014; Greene, 2015). Many different brain areas are implicated in making moral judgements, yet these areas are disparate and carry out many different functions (Fumagalli and Priori, 2012). Individually, these brain regions are involved in a broad range of processes, from emotional valence (the positive or negative nature of emotional arousal) to conscious deliberation and "theory of mind", which is the ability to attribute mental states and beliefs to others.

The neuroscience of moral judgement has an extensive history. Initially it was primarily a study of lesions and brain injuries, which appeared to have led to personality changes, abnormal social conduct and functional loss. In the last two decades, neuroimaging techniques such as EEG, fMRI and TMS have become increasingly prevalent in cognitive neuroscience. Naturally these techniques were introduced into the study of moral cognition, and in the seminal study by Greene et al. (2001), fMRI was used for the first time in the identification of the neural correlates of moral judgements. Recently, the field of moral cognition has greatly expanded, and morality-related words saw an eight-fold increase in the literature (Greene, 2015).

Put broadly, moral judgements are judgements about whether particular actions, entities or norms are "right or wrong", or "good or bad". As discussed briefly in the introduction, it is a difficult task to decide what actions and entities count as being morally relevant. However, it is necessary to posit a specific definition so that the

phenomenon is suitable for neuroscientific inquiry. Challenges in conceptualising moral judgements and more broadly, moral cognition, will be discussed in chapter two. For now, I will briefly address some conventional aspects of "moral judgement" researchers have focused on in this field.

A common use of the term moral judgement pertains to evaluating actions resulting in the harming of other individuals. In an example of one such study, Greene et al. (2001) presented participants with "trolley problems", including the "switch" and the "footbridge" variants[2], to examine what brain regions become active when participants made judgements in response to an impersonal dilemma (the "switch variant") or a more personal dilemma (the "footbridge" variant). The example above is centrally concerned with moral judgements involving harm, specifically actively harming one individual to prevent harm to a greater number of others.

Studying judgements concerning harmful actions is just one potential avenue to explore in moral cognition. Other commonly studied aspects involve evaluating actions based on fairness or justice (Decety and Yoder, 2016). More generally some studies assess "moral emotions", which are reactions to a variety of morally salient stimuli, such as war scenes, poor children and assaults (Moll et al., 2002). Because of the diversity, it seems obvious then that moral judgement in humans would involve many different functional aspects of the brain, from the emotive to the cognitive, which will be explored now.

1.1 The Moral Network

In this section I will summarise and explore some of the regions commonly implicated in moral judgement. I will reference a few select studies, with a diverse set of methodologies, that exemplify what is known about the function and roles of these brain regions within the moral network. The implicated brain regions are well

[2] Trolley problems, as a thought experiment were initially formulated by Foot in Foot, P. 1978. Virtues and Vices and Others Essays in Moral Philosophy. Berkeley: University of California Press. And further developed by Thomson in Thomson, J. J. (1985) 'The Trolley Problem', *The Yale Law Journal*, 94(6), pp. 1395-1415. Trolley dilemmas are a form of sacrificial dilemma, where a runaway trolley is endangering several individuals tied the upcoming track. One must decide whether to make a sacrifice to save those individuals. In the 'switch' variant, one can pull a lever to divert the trolley away from five individuals and onto a track with only one individual. In the 'footbridge' variant, the observer is standing on a footbridge above the track accompanied by another large individual. The decision here is whether one would push the large individual off the footbridge and onto the track thus stopping the trolley, to save five individuals.

documented through the extensive work done in cognitive neuroscience over the past few decades. Both cortical and subcortical structures are involved.[3]

I will focus primarily on four important regions, as these will be consistently referred to in the remainder of the thesis. These are the amygdala, the ventromedial prefrontal cortex, the dorsolateral prefrontal cortex, and the temporoparietal junction. There are many other brain regions implicated in moral cognition, and I will briefly review to a select few of them, but they are of lesser importance to the overall discussion in this thesis. The objectives of this section are to summarise the functioning of these brain regions in moral cognition and to describe the conventional form of studies in the field. Finally, through a conventional review of the neuroscientific literature I will draw some general conclusions about the nature of moral cognition.

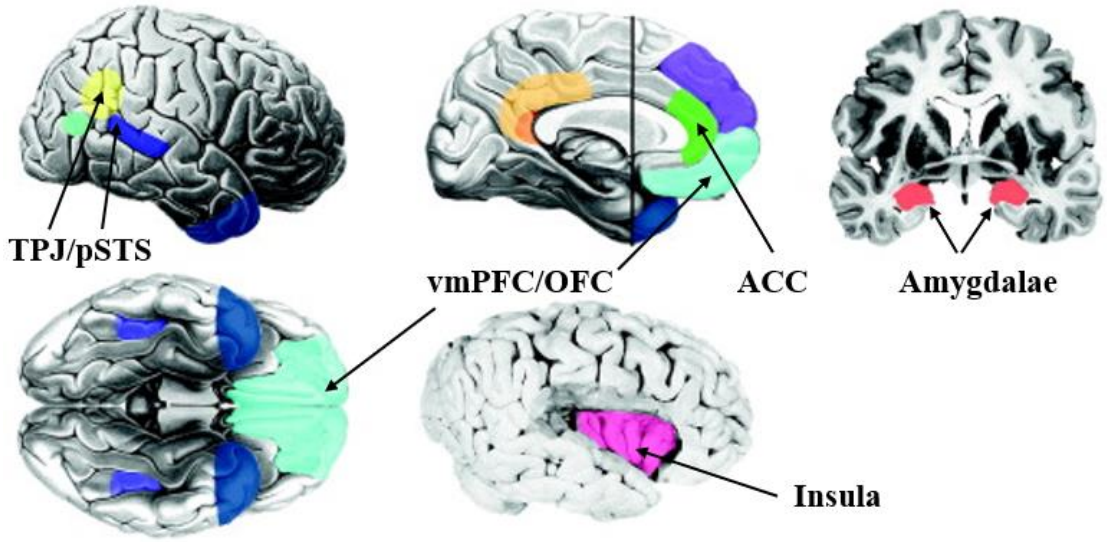

Figure 1 The human brain with regions relevant to moral cognition highlighted. Relevant regions are labelled. The amygdala, represented by the red, is a structure deep in the brain, towards the inferior surface. The Ventromedial prefrontal cortex (vmPFC), represented by the teal, is on the inferior surface of the prefrontal cortex and extends medially. The Anterior Cingulate Cortex (ACC), represented by the green, is a midline structure situated anteriorly. The Insula, represented by the pink, is a lobe of the brain that is deep to the brains lateral surface. The Temporoparietal junction and the posterior superior temporal sulcus (TPJ/pSTS), represented by the yellow and blue respectively, are closely situated structures, positioned at the margin of the temporal and parietal lobes. This figure is modified from Kennedy and Adolphs (2012).

[3] There are several recent reviews which examine these brain structures, which can be referred to if more detail is required: Mendez, M. F. (2009) 'The neurobiology of moral behavior: Review and neuropsychiatric implications', *CNS Spectrums*, 14(11), pp. 608-620.; Fumagalli, M. and Priori, A. (2012) 'Functional and clinical neuroanatomy of morality', *Brain: A Journal of Neurology*, 135(7), pp. 2006-2021.

1.1.1 The Amygdala

The amygdalae (plural of amygdala) are two nuclei, one right and one left, located deep in each temporal lobe, anterior to the hippocampus. This structure has shown to be involved in attention, affective arousal and emotional valence (Lane et al., 1999).

In a study by Moll et al. (2002), participants underwent fMRI, while observing non-moral unpleasant scenes or pictures (such as dangerous animals and bodily fluids), as well as observation morally objectionable scenes and pictures (abandoned children and war scenes). Relative to neutral pictures, amygdala activation was observed in both conditions. This suggests that the amygdala is a common neural substrate in the processing of both moral and basic unpleasant emotions. A similar relationship is observed with the upper midbrain bilaterally, periaqueductal grey matter, right thalamus and superior colliculus, right insula/inferior frontal gyrus, bilateral posterior temporal–occipital cortex and right intraparietal sulcus, all showing activation, which is suggestive of a common neural substrate in the processing of unpleasant emotions with both moral and non-moral dimensions (Moll et al., 2002). Therefore the amygdala does not seem to be a module underlying specific functions, but rather is part of a network in each case.

The specific role of the amygdala in moral cognition was explored further by Shenhav and Greene (2014). In this study participants were presented with moral dilemmas similar to the classic "trolley problems", where one individual agent could act to save a greater number of lives by actively harming another individual, or could refuse to harm leading to a greater number of deaths.[4] Subsequently participants were asked which of the two potential responses to the dilemma would produce "better results" (utility assessment)[5], which the participant would feel worse about doing (emotional assessment) and finally which do they find more morally acceptable (integrative moral judgement). Brain activity associated with performing the tasks was measured using fMRI. Amygdala activity was correlated with negative ratings for how the "utilitarian"[6] response would feel, i.e. how emotionally aversive that choice was. Amygdala activity was also found to be negatively correlated with the "utilitarian"

[4] An example of such a dilemma can be seen in Figure 1 of the referenced paper: Shenhav, A. and Greene, J. D. (2014) 'Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex', *The Journal of Neuroscience*, 34(13), pp. 4741-4749.
[5] Utility in this case refers to something of value, e.g. lives, pleasure, happiness.
[6] The response to the moral dilemma opting to cause harm to save more lives.

response when participants made the "all things considered" (integrative) moral judgement. Coupling of the amygdala and the ventromedial prefrontal cortex (vmPFC) was greatest during the emotional assessment and the lowest during the utility assessment. These findings are consistent with the claim that the amygdala plays an important role in moral cognition, specifically in the emotional assessment of salient stimuli, in response to potentially harmful actions. Shenhav and Greene (2014) have hypothesised that their results indicate that the amygdala can guide moral behaviour through its connectivity with the vmPFC. In summary, the evidence suggests that the amygdala has a role in normal moral and social processing.

1.1.2 Ventromedial Prefrontal Cortex

The Ventromedial prefrontal cortex (vmPFC) is part of the prefrontal cortex (which is involved in higher cognitive processes) and is located just above the position of the eyes, on the inferior surface of the brain. The vmPFC is commonly associated and physically overlaps with the orbital frontal cortex (OFC), a region similarly on the inferior surface of the frontal lobe. These two regions will be discussed together. The vmPFC is an area commonly implicated in social reasoning and decision making.

Patients with damage to the vmPFC exhibit abnormal social conduct. The most famous example of such as patient was Phineas Gage, who provides an early case of brain trauma being linked to specific social and moral abnormalities. Gage was a railroad construction foreman in New England in the mid-1800s. On the 13th of September 1848, Gage was injured in an accident while blasting rocks to make way for new railway. An explosion caused a tamping iron[7] to pass through his skull, destroying part of his brain. Amazingly, despite the extent of his injury, Gage remained conscious and survived (Harlow, 1999).

Gage's physician Henry Harlowe described his patient's post-accident changes in this way:

> He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious and vacillating, devising many

[7] A large iron rod used to compact sand on top of ignition powder. Gage became distracted while using a tamping iron, leading to ignition of the powder.

> plans of future operation, which are no sooner arranged than they are abandoned in turn for others appearing more feasible.
>
> (Harlow, 1993)

Famously, Harlowe (1993) reported that many of Gage's friends and family said he was "no longer Gage".

A famous study by Damasio et al. (1994) used medical imaging on Gage's skull to calculate which brain areas would have been affected. They concluded the vmPFC was most likely damaged, and that Gage's reported behavioural changes are like those of modern patients with vmPFC lesions. However, Macmillan and Lena (2010) have argued that Gage's case is commonly exaggerated, in that many researches retrospectively embellish Gage's story and behavioural changes, whereas in reality Gage recovered and adapted psychologically to his injury. Nevertheless, the neurologist Antonio Damasio studied similar patients with vmPFC lesions and noted changes in their social and moral behaviour, such as an increased propensity to make risky social and financial decisions (Damasio et al., 1990), which indicates there is some credibility to the link that Gage is used to illustrate. Damasio (1996) developed the "somatic marker hypothesis" to explain these abnormalities. This hypothesis argues that "somatic markers" (feelings that are embodied) are important in guiding decision-making and practical reasoning, and have an influence through the vmPFC, which if damaged severs the link between feelings and reasoning.

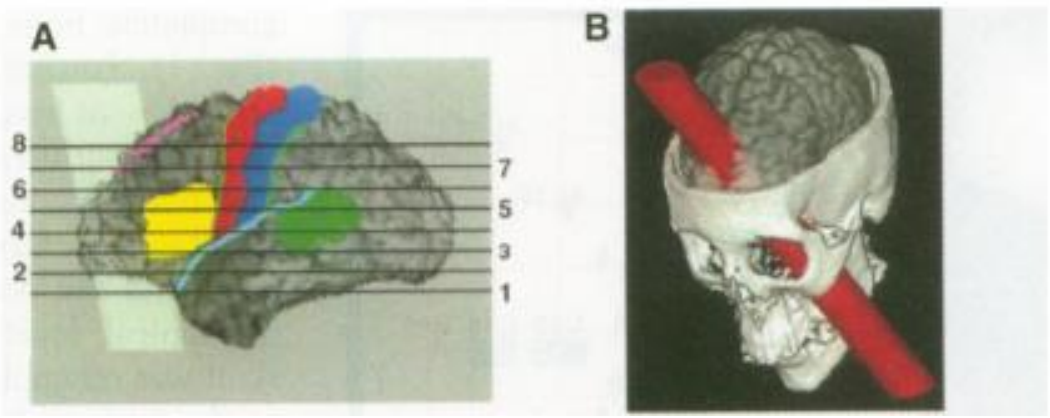

Figure 2 The brain and skull of Phineas Gage. (**A**) A lateral brain view. The tamping iron, represented in solid white, is shown in its estimated position in relation to Phineas Gage's brain. (**B**) The tamping iron, represented in solid red, is shown in its estimated position relative to Phineas Gage's brain and skull. The bilateral vmPFC is estimated to have been lesioned. The figure is modified from Damasio et al. (1994).

In the earlier mentioned study by Moll et al. (2002) brain regions that activate when showing participants moral unpleasant images compared to non-moral unpleasant images were also examined. Three regions showed relatively increased activation to the morally salient stimuli. These regions were the right posterior superior temporal sulcus (STS) the medial frontal gyrus (MedFG) and the right OFC (a region closely associated with the vmPFC). This suggests the importance of the vmPFC in processing moral scenarios that elicit emotions. Similarly, the vmPFC has been consistently implicated in automatic moral judgements in response to emotionally laden dilemmas (Greene et al., 2001; Greene et al., 2004).

In a review by Padoa-Schioppa and Cai (2011) that focused on the OFC (vmPFC), the researchers collated findings that implicated this structure in economic decision making. In response to this literature these researchers proposed that the general functioning of the OFC is to compute subjective values for abstract representations of goods that are behaviourally relevant (Padoa-Schioppa and Cai, 2011). It has been shown that the vmPFC carries out a similar function in moral judgement, with the vmPFC being sensitive to the "moral value" of specific actions and outcomes (Shenhav and Greene, 2010). These findings reflect the idea that the moral network (the network of brain regions implicated in responding to moral judgements) involves brain regions that perform domain general functions, which are generalisable and foundational functions that can be translated to operate in many domains, such as the "moral domain".

As discussed above, in the study by Shenhav and Greene (2014), coupling of the amygdala and the vmPFC was greatest during the emotional assessment of stimuli, and lowest during the assessment of utility. They also found that vmPFC activity was greatest during the "all things considered" moral judgement, where after being exposed to a moral dilemma and the two potential response options the participants were asked "which do you find more morally acceptable". This judgement was assumed to involve both affective and cognitive processes. They concluded that the vmPFC played a role in integrating an affective signal from the amygdala with a "utilitarian" assessment from other brain structures (Shenhav and Greene, 2014).

Damage to the vmPFC has been associated with an increase in "utilitarian" judgements when individuals are faced with moral dilemmas (Koenigs et al., 2007). In

response to their own study's findings Shenhav and Greene (2014) hypothesised that "utilitarian" decision making can still occur independent of the vmPFC; however, paradoxically, the amygdala relies on the vmPFC to exert control on behaviour. This leads to the proposal that when the vmPFC is damaged, an increase in "utilitarian" judgements are observed; not because of damage to a structure involved with an intrinsically affective process, but rather due to damage of the conduit that integrates affective processes into choices associated with decision making (Shenhav and Greene, 2014).

1.1.3 Dorsolateral Prefrontal Cortex

The dorsolateral prefrontal cortex (dlPFC) is another specific area within the prefrontal cortex and is located on its superior and lateral surface, inferior to the upper left and right margins of the forehead. The dlPFC is implicated in executive control and famously is one of the final areas of the brain to fully develop through adolescence and young adulthood (Sowell et al., 2001).

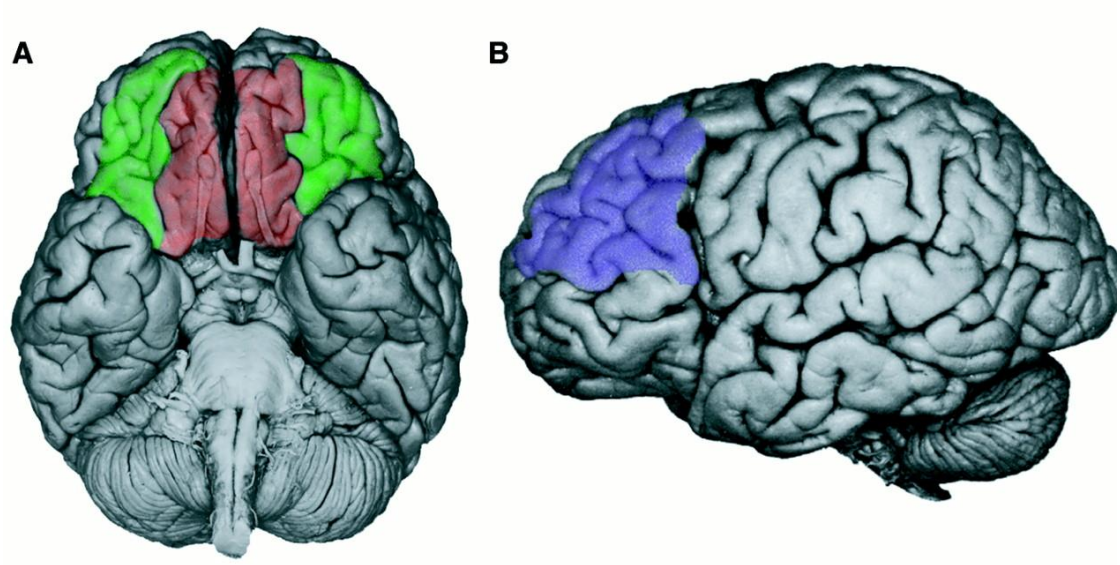

Figure 3 The ventromedial prefrontal cortex and the dorsolateral prefrontal cortex. (**A**) An inferior view of the brain (looking from the bottom up). This figure shows the interelation of the vmPFC (seen in red) and the OFC (seen in the green and red). Note that the vmPFC actually extends into the midline of the brain, as seen in **Figure 1**. (**B**) A lateral view of the brain (looking from the side). This figure shows the dorsolateral prefrontal cortex (dlPFC), which is situated on the upper outer portion of the prefrontal cortex. The figure is modified from Davidson et al. (2000).

An interesting study by Steinbeis et al. (2016) illustrates the general functioning of the dlPFC. This study looked at decision making in children (6 to 13 years old),

examining their preference between short and long-term rewards, with the long-term rewards being of greater magnitude. A preference for the more immediate reward is called temporal discounting. It was shown that older children were more likely to choose the greater long-term reward compared to younger children. This could either be explained by the subjective values of long-term rewards increasing over time, as children get older, or by greater "top-down" control in older children who resist the temptation towards short-term rewards, or both. To test these hypotheses, using fMRI, Steinbeis and colleagues examined the functional connectivity between the vmPFC (which, as mentioned previously computes subjective values) and areas responsible for executive control such as the dlPFC. They found that the increasing preference for long-term rewards was not related to any difference of choice-independent valuation. Rather, it correlated with greater vmPFC – left dlPFC connectivity. The researchers concluded that as people age, and the dlPFC matures individuals are better able to exhibit behavioural control and forego immediate pleasures (Steinbeis et al., 2016).

In the moral realm the dlPFC has been shown to exhibit a similar function as described above. Greene at al. (2004) showed that the dlPFC was more active during processing of impersonal moral dilemmas associated with a participant indirectly causing harm to save others, and less active during personal moral dilemmas, where the harm is more direct and the participant is more responsible. More importantly, the dlPFC showed greater activation in the participants that chose the option that most maximised utility, when faced with the personal "trolley-like" dilemmas. They concluded that the dlPFC was executing control to regulate emotions that opposed causing harm to the sacrificial individual, allowing the action that maximised utility to be performed. On top of this, they concluded that the dlPFC has a more direct role on the outcome of the moral judgement, as it plays a role in abstract reasoning, such as assessing utility (Greene et al., 2004).

One final study shows the importance of the dlPFC in the application of rule-based behaviour. Prehn et al. (2007) studied which brain areas became activated when a participant made a socio-normative judgement (such as whether it's OK to smash glass on a train), as a function of the participant's moral competence. Moral competence was defined as the participant's ability to "apply moral orientations and principles in a consistent and differentiated manner in varying social situations" (Prehn et al., 2007), and was measured using the MJT (Moral Judgement Test) taken from Lind (2008). The

researchers found that activity in the dlPFC was inversely related to moral competence and was more active, on average, when participants with low moral competence made socio-normative judgements. The researchers proposed that the dlPFC is important in people's conscious execution of rule-based behaviour as a compensatory mechanism for a failure to consistently apply moral principles automatically. The dlPFC is thus an important structure in abstract cognition in moral judgements, specifically regarding rule-based behaviour and executive control.

1.1.4 Right Temporoparietal Junction/Posterior Superior Temporal Sulcus

These two regions are close structurally and many papers refer to them together or single out one while referring to the brain region in general. Hence, the two structures will be discussed together. They are the temporoparietal junction (TPJ), more specifically the right TPJ, and the posterior superior temporal sulcus (PSTS). These two regions are roughly located deep to and superior to the ear relative to the skull. The TPJ is the literal junction between the parietal and the temporal lobes. The PSTS is consistently implicated in the moral network and is one of the three regions identified by Moll et al. (2002) that is active specifically when participants view morally unpleasant images, compared to non-moral unpleasant images.

The TPJ has been shown to have a variety of different roles. A meta-analysis of the neuroimaging literature was conducted by Decety and Lamm (2007). According to the review the TPJ has been implicated in many complicated cognitive functions, such as theory of mind, empathy (sharing emotions with others), agency (the attribution of actions to oneself or another) and re-orientating attention (switching attention between different available/significant stimuli).

Let us focus on theory of mind (ToM), which the TPJ and the PSTS have both been implicated in (Saxe and Wexler, 2005). ToM has obvious importance in general social functioning and decision making, however it also plays a role in moral cognition. Young et al. (2010) conducted an experiment using transcranial magnetic stimulation (TMS) to temporarily disrupt the neural activity in the right TPJ in participants and observed the effect on moral judgements. The participants were split into control (where the TMS was used 5cm behind the TPJ) and TPJ-TMS groups, and both completed a series of moral judgement tasks with four variants, concerning the

intentions of characters and outcomes of different potentially harmful events. The TPJ disruption group judged less harshly than the control group characters who attempted to harm yet failed to cause harm to another individual. Young et al. (2010) concluded that the TPJ is involved in belief attribution which is essential for making typical moral judgements. The disruption of the TPJ led to a misalignment between the intentions of a malicious actor and the participant's assessment of that actor, with the participant's assessment relying more on the consequences of the action.

Another study, by Decety and Cacioppo (2012) used high-density electrical neuroimaging to examine the patterns of connection between the PSTS/TPJ, the amygdala and the vmPFC when participants viewed videos depicting someone causing either intentional harm or accidental harm to another person. High-density ERPs were used to determine the spatio-temporal dynamics between those regions, i.e. the specific times they became active relative to each other. The researchers found that the first region to become active was the PSTS, followed by the amygdala (emotional processing), and then the vmPFC. The results are consistent with the view that the PSTS/TPJ are involved in the processing of intentionality that is necessary for these kinds of moral judgements and their activity precedes the judgement itself (Decety and Cacioppo, 2012). Taken together, these studies illustrate the importance of the PSTS/TPJ in social cognition, and how early categorisation, such as belief attribution, directs our moral judgements, before emotional inputs, executive control and abstract reasoning integrate into the judgement.

The TPJ has also been specifically implicated in processing difficult moral decisions. In a study by Feldmanhall et al. (2014), researchers used fMRI to examine the differences in brain activity when people are exposed to either "easy" or "difficult" moral dilemmas. Difficult moral decisions were defined as responses to moral dilemmas where there is no obvious consensus amongst the participants (FeldmanHall et al., 2014). Subsequently, to determine whether their distinction between easy and difficult moral decisions was accurate, the researchers asked participants after the experiment to rate the difficulty of the dilemmas they were given. Participants consistently rated the preordained "difficult" moral dilemmas as more difficult and vice versa for the "easy" dilemmas. During difficult moral dilemmas it was found that the vmPFC decreases in activity while activity in the TPJ and regions inferiorly along the temporal lobes increases. The inverse was found during easy moral decisions.

Importantly the effects were not observed during difficult non-moral dilemmas, so this relationship is specific for moral dilemmas. Feldmanhall et al. (2014) propose that the reliance on the TPJ during difficult moral decisions reflects the use of a more reflective cognitive network that involves attentional shifting allowing for the weighing up of relevant cognitive and social stimuli.

The studies discussed above reveal the extremely important role that the TPJ/PSTS plays in moral cognition and social processing more generally. It has been shown to be a primary brain region involved in theory of mind – which some theorists refer to as "cognitive empathy", the ability to understand what another person is thinking[8] - as well as shifting attention to different stimuli, which may point to an ability to reflect upon different reasons before forming a judgement.

1.1.5 Other structures: Anterior Cingulate Cortex and Insula

The anterior cingulate cortex (ACC) is a midline structure towards the anterior (the front) of the brain, but deep to the prefrontal cortex that was discussed earlier. Alongside the amygdala it is referred to part of the "limbic" system which is associated with emotional processing. Like the TPJ/PSTS the ACC has been implicated in a variety of different functions.

Using fMRI Singer et al. (2004) showed that the ACC was activated when participants experienced pain, and when they perceived their loved ones experience the same pain. The response of the ACC also correlated with the empathy scores of the participant. These results illustrate the role that ACC plays in empathy, as it is involved in processing the affective component of pain both first hand and empathically. In the same study a similar relationship was found with the anterior insula. This form of empathy is often referred to as "emotional empathy" – where one feels what one perceives someone else is feeling (Bloom, 2016). Empathy for others, especially regarding physical pain is important for moral processing. It alerts individuals to the suffering of others, makes the suffering salient and is a motivator.

The ACC also plays an important role in cognitive conflict resolution. In the study by Greene at al. (2004), the ACC alongside the dlPFC showed greater activation during "difficult" moral dilemmas. Greene et al. (2004) hypothesised that the ACC is

[8] Take for example Paul Bloom: Bloom, P. (2016) *Against empathy: The case for rational compassion.* The Bodley Head.

highly active in detecting conflicts and resolving conflicts through a recruitment of the dlPFC.

Like the ACC, the insula is another deep brain structure involved in the limbic system and consists of two lobes. The insula lobes are located either side of the corpus striatum (the anatomical core of the brain). The insula is typically related to emotional processing. The study by Moll et al. (2002) discussed earlier identified insula activation in participants while they observed both morally unpleasant and non-moral unpleasant images. As mentioned in the ACC section, the insula is involved in empathy. In the aforementioned study by Singer et al. (2004), the insula was active when participants experienced pain and when they observed their loved ones experience the same pain.

The anterior insula also plays a role in the emotion of disgust. In a study by Wicker et al. (2003), fMRI was used to measure levels of insula activity when participants smelled a disgusting odorant, and when they viewed the face of an actor smelling and odour and responding with disgust. In both cases the activity in the anterior insula increased. The previous two studies suggest that the activity in the brain of the empathisers is similar to the activity in the brain of the person they are observing. Although this form of empathy, emotional empathy, is also called "affect sharing", the empathic emotion is actually induced in the empathiser (Singer, 2006). Despite being similar, the empathic emotion and the genuine emotion are not indistinguishable (Singer, 2006).

Emotional processing in general, in the insula, as well as the ACC and amygdala is integral to moral cognition. It has been shown that emotions like disgust can influence and bias moral judgements. In a study by Olatunji et al. (2016) participants were split into three groups. One group had their hands submerged in fake vomit to elicit disgust, another had their hands submerged in cold water to elicit discomfort and the third group was a control. Each group were given a set of scenarios containing moral violations and were told to rate how morally wrong those violations were. The disgust group rated purity violations (e.g. "John spit in someone's drink" and "John knowingly served someone food past its expiration date") as being morally worse than the control and discomfort groups.

1.2 Features of the Moral Network

All six of these brain regions discussed in this chapter have been consistently implicated in moral judgements, in both lesion and neuroimaging studies. Despite their disparate locations in the brain, these brain areas are linked at a higher functional level, in the process of moral cognition, where each area plays a specific role within the brain's "moral network".

Each brain region discussed is also implicated in other functions, outside of moral cognition. This is because each region has domain-general functions, meaning that they perform more general functions that play a role in many different higher-cognitive processes. For example, the vmPFC is implicated in economic decision making as well as moral cognition, with a domain-general function underlying both processes. As discussed earlier, the structure has been theorised to compute subjective values for abstract representations of specific goods that are behaviourally relevant (Padoa-Schioppa and Cai, 2011). In both moral and economic contexts this function is important. Making moral judgements draws upon the neural circuitry associated with this structure, to compute the "moral" value of actions or outcomes (Shenhav and Greene, 2010). Similarly, the anterior insula is implicated in domain-general functions which translated to higher-level functions. This structure processes general emotions such as disgust (Wicker et al., 2003), which can then be involved in moral cognition (Moll et al., 2002). Finally, the TPJ has been implicated in theory of mind and attention shifting (Saxe and Wexler, 2005), and has also been shown to play an important role in moral cognition, especially during difficult moral dilemmas when one is weighing up many different socially relevant stimuli (FeldmanHall et al., 2014). Thus, there is no specific brain structure solely responsible for moral function.

In summary, the findings of recent neurological and psychological research into moral cognition suggest that moral cognition is a diverse process. It consists of regions that are implicated in multiple cognitive processes such as exercising executive control (dlPFC), attention switching between stimuli (TPJ/PSTS) and computing subjective values (vmPFC). Affective regions are involved that have been shown to bias (insula), inform through empathy (insula and ACC), and directly integrate into processing of moral judgements (amygdala through the vmPFC). Regions involved with social processes also play an integral role in moral cognition. Processes like theory of

mind/cognitive empathy are important in belief attribution (TPJ/PSTS), and emotional empathy plays a role in moral motivation and raising awareness of others in need (insula and ACC). Any psychological theory of moral cognition must contend with the evidence of its plurality.

1.3 Conclusion

In this chapter I have discussed selected contemporary literature on the neuroscience of moral judgement, with a particular focus on the key brain regions implicated in a "moral network" and the complex faculty of empathy. I have shown that moral cognition is a complex process involving the integration of these brain regions, which are disparate in space and function. These brain regions build off their domain-general functions to carry out the higher-level process of moral cognition. These findings suggest a multi-faceted and integrative model of moral cognition. By integrative, I mean that the implicated brain regions function as a network, where there is intense connectivity and associations between them. By multi-faceted, I mean that the implicated brain regions and their functions are diverse, involving processes related to emotion, reasoning and social cognition. This conclusion however requires more support and is contrary to several prominent mainstream accounts of moral cognition. This neuroscientifically integrated feature is part of an "integrative account of moral cognition" and will receive greater focus in chapters two and four of this thesis.

In chapter two I will set out three major problems facing the neuroscience of moral judgement. These challenges will set out where care needs to be taken when using the results of empirical neuroscience, and how when used alone, they inadequately describe human moral cognition. Despite these challenges, many important findings and results have come out of cognitive neuroscience, referenced by the discussion and conclusions of this chapter. In chapter three I will introduce a neo-Kantian account of moral obligation and in chapter four, I will relate this moral philosophy to current neuroscience. This work will aid in the development of an integrative account of moral cognition.

2 Challenges with the Cognitive Neuroscience of Morality

The neuroscientific study of morality requires an integrative approach. Two kinds of integration are required. The first relates to methodological integration in the form of an interdisciplinary analysis. The second relates to the integration of a brain network: it concerns the need for an account that explains how moral cognition results from various integrated brain processes. The need for such an account was discussed in chapter one, where I showed how key neuroscientific studies of moral cognition suggest it to be a multi-faceted process, involving emotion, reason and social cognition. Brain regions disparate in space and function, operate as a network, that "builds off" each regions domain general functions.

In this chapter I will explain why an interdisciplinary analysis is needed, by outlining the methodological challenges facing the current neuroscience of moral cognition. These challenges can be clustered into three groups. In each grouping I will discuss the challenges in relation to both the underlying empirical neuroscience and theories in moral philosophy.

The first set of challenges concern the relationship between moral philosophy and models of moral cognition, which attempt to systematically describe principles and processes common to moral thinking. Many neuroscientific models of moral cognition have roots in older ethics theories. There is a common assumption in the literature that these traditional theories, such as utilitarianism, deontology and virtue ethics are represented in neural circuitry. There are methodological problems with this neuroscientific assumption. For example, it is difficult if not impossible to accurately attribute certain kinds of ethical reasoning to a participant based on their responses to moral dilemmas in an experiment. What is needed is a more nuanced use of moral philosophy, that moves away from simplistic and traditional versions of theories and towards versions that are neuroscientifically informed.

The second category are "circularity" challenges. These involve challenges in conceptualising moral judgements, crafting methodologies to study them, interpreting any findings and relating models of moral cognition to these findings. There is a potentially problematic circularity that underlies this process. To explore moral

cognition, one must adopt a working definition of "morality" and then "moral cognition". These conceptualisations in turn influence what cognitive faculties are considered as relevant to moral cognition and thus worth exploring. This in turn influences the kinds of methodologies employed to study these relevant faculties and the interpretation of any findings.

The third category of challenges concern the methodological limits in the neuroscience of moral judgement. Due to technological limitations in current experimental techniques, cognitive neuroscience cannot capture how the brain regions involved in the moral network integrate as the "higher-function" of moral cognition. Secondly, most neuroimaging studies only capture up to a few seconds of brain activity around the making of a moral judgement.[9] This fails to capture the complexity of moral judgments which extend much further temporally, including minutes, hours and even days. Furthermore, many experiments in neuroscience fail to capture the real-life conditions and context that moral judgements are made in. This means that they lack ecologically validity, and hence their results do not accurately describe moral cognition.

In the following sections of this chapter I will discuss these problems in more detail and propose some potential solutions. Some of these challenges can be overcome and the neuroscience can be improved. Ultimately though, some deep problems remain. This shows the need for an interdisciplinary approach to the study of moral cognition that utilises empirical neuroscience alongside social psychology, moral philosophy and theoretical neuroscience. In particular, neuroscientific study depends on moral philosophy to conceptualise moral cognition and interpret any findings that emerge.

2.1 The Relationship Between Moral Philosophy and Models of Moral Cognition

Alongside the recent empirical search for the brain regions and processes involved in moral judgement, models of moral cognition have been developed. The philosophical inquiry into the nature of human moral thinking, however, is far older and in many ways anticipates the current debates. In this section I will lay out the prominent historical and modern models of moral judgement. The models will be described using

[9] In EEG specifically, this time window is even less, with only a few hundred milliseconds of brain activity being recorded.

two polarities: the first concerns the roles of reason and emotion, and the second concerns the level of dissociation or integration of the processes involved in moral cognition. There is an observable and pervasive link between moral philosophy and these neuroscientific models of moral cognition. While moral philosophy is needed to inform neuroscience and the interpretation of any findings, the link may not be so direct as many theorists have posited. There is a popular assumption in the literature that normative ethics theories should be represented in neural circuits. Often, simplified and historical theories receive the most focus. This assumption is methodologically dubious, as it can be difficult to prove that certain responses participants give to moral dilemmas in fact reflect processing akin to a specific normative ethical theory. What is needed is a more nuanced use of morally philosophy, that moves away from simplistic and historical versions of theories and towards versions that are neuroscientifically informed.

2.1.1 Rationalism and Sentimentalism

One long-standing philosophical debate currently being reiterated in the psychological and neuroscientific study of moral cognition is between moral rationalism and moral sentimentalism. Moral rationalism, often associated with Kant, asserts that moral thinking is fundamentally a rational process, and that moral truths can be known *a priori*. By contrast, moral sentimentalism, often associated with David Hume, asserts that moral truths are grounded in human emotional responses to experiences. Several important modern models of moral cognition have roots in these older philosophical theories, and interpret the empirical evidence with the conceptual framework they provide. The contemporary application of these theories is often expressed through a focus on roles of reason and emotion. This focus concerns whether these two faculties are inputs or outputs of the moral judgement, and whether they constitute or merely inform the moral judgement (Prinz, 2016).

Modern rationalist models assert that pure reason, practical reason, deliberation and cognitive control play the most significant role in the making of moral judgements. These models do not deny the association of emotions with morality, but hold that they play a less important role. For example, one might hold that stealing is morally wrong on the basis of certain reasons, and then after observing an act of stealing become angry, disgusted or saddened. However, even if these feelings are based on the moral

judgement, the judgement itself is based on the reasoning. Immanuel Kant is a famous historical proponent of rationalist theory. The moral law, he argued, can be distilled into the following "categorical imperative": "Act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 2002). According to Kant, the evaluation of moral actions is therefore purely based on the requirements of reason.

The primary "modern" advocate of rationalist theorists in the field of psychology was Lawrence Kohlberg. Kohlberg was highly influential in the 20th century, before the emergence of the cognitive neuroscience of moral judgement, with its extensive use of neuroimaging. Therefore, despite being the most notable proponent of rationalist theories in moral psychology, his theories are largely historical. Kohlberg (1971) proposed a six staged theory of moral development, and with each stage: "…the obligation to preserve human life becomes more categorical, more independent of the aims of the actor, of the commands or opinions of others…"

Sentimentalist models of moral cognition assert that emotions and sentiments are the major constituents of moral judgements. Hence, such theories are the opposite of rationalist theories. Importantly, sentimentalist models do not deny any role for reasoning in morality, however they argue that it is subservient to emotions. David Hume is a famous historical proponent of a sentimentalist theory of moral cognition. He argued that reason is only ever a "slave of the passions", and that moral judgements are constituted by feelings of approbation and disapprobation, i.e. praise and blame respectively (Hume, 2007, p. 266). Jesse Prinz is a modern proponent of sentimentalist theories, and he argues that sentiments, which are dispositions to certain emotions, underlie moral judgements (Prinz, 2011). These important emotions are anger and disgust when one observes an immoral act, and guilt and shame when one performs an immoral act. Prinz argues that an action is wrong if one experiences a sentiment for these emotions when either observing or performing that action (Prinz, 2011).

In the early 2000's, following the publication of a famous article by Haidt (2001) "The Emotional Dog and Rational Tail" there was a major decline in the popularity of rationalist theories within neuroscience. In this article Haidt proposed the "social intuitionist model" of moral judgement. In this model emotions and intuitions lead to moral judgements, and reasoning is a post-hoc process, meaning that it is

performed after the judgement is made (Haidt, 2001). Haidt (2001) admits that reasoning may lead to a moral judgement, but this is very rarely the case.

In summary, modern models of moral cognition often involve a contemporary application of older theories from moral philosophy. A traditional philosophical conflict between rationalism and sentimentalism is particularly noticeable, in the treatment of emotion and reason as oppositional faculties.

2.1.2 Integration and Dissociation

Modern neuroscience has made extreme versions of sentimentalism and rationalism untenable, and so most of the current prominent models are hybridised to some extent. Hybridised models of moral cognition assert that both reason and emotion play important roles in the making of moral judgements. However, these models are highly diverse, with both reason and emotion, among other mental faculties potentially playing multiple roles that differ greatly between different models. One such variability concerns the level of integration of the various neural components in moral cognition. On one end, the neural components are highly dissociated, compete against each other and operate independently. On the other side the neural components are highly integrated and operate as a network. The components may involve emotional and rational processes.

Greene's "dual process model" of moral judgement is an example of a dissociative hybrid model (Greene, 2016). This theory holds that the brain has two avenues for making moral judgements. The first is automatic, intuitive and emotionally informed and often leads to what Greene (2016) categorises as "deontological judgements", i.e. judgements that deem the action itself is judged as right or wrong irrespective of its consequences. The second is deliberative and involves the cognitive control of emotions. Greene (2016) claims that this leads to "utilitarian judgements", where the value of an action is calculated based on its consequences. According to his "dual process model", moral cognition is highly dissociated, as emotion and reasoning are involved in two separable modes of making judgements, these being automatic and deliberative.

The model proposed by Casebeer and Churchland (2003) transcends the traditional dissociation between reason and emotion. Their model is best represented by a neo-Aristotelian virtue theory, where morality concerns what one should do and think

to function well as a human. They argue that the diverse brain processes involved in moral cognition are collaborative. Emotive and cognitive processes integrate with empathy, perspective taking, and one's memory system to allow an organism to "think about and actually behave in a manner enabling it to function as best it can" (Casebeer and Churchland, 2003). Moll et al. (2008) advocate for a form of hybridised sentimentalism, where emotion and reason both contribute to altruistic motivations, which they claim, underlie moral actions. They claim that processes involved in reason and emotion are nondissociable, and are represented in integrated cortico-limbic neural assemblies, that constitute such moral motivations. Moll et al. and Casebeer and Churchland, hold that moral cognition involves the collaboration and integration of processes that are both emotive and rational (Casebeer and Churchland, 2003; Moll et al., 2008). Thus, compared to Greene's dual process model, these two models represent significant alternatives to the typical reason-emotion dichotomy.

2.1.3 Should We Expect Moral Theories to be Reflected in Neurological Processes?

Another broad challenge that arises in the testing of moral cognition is whether we should expect cognitive or neurological processes to reflect theories in moral philosophy. There is a common assumption in the literature that competing normative ethics theories, such as deontology, utilitarianism and virtue theory will be represented in different neural networks and brain processes. In other words, there is an idea that normative theories would either correspond with or manifest from a particular brain region or network of regions.

The assumption that multiple normative ethics theories are represented in different neural networks involves the idea that these neural networks mirror the functioning of the theories they represent and compete with each other, much like the theories do in moral philosophy. Greene's Dual Process theory of moral judgement, discussed earlier in this chapter, is a prominent case that illustrates this assumption. Remember, according to this theory, when someone is given a moral dilemma, depending on the specifics of the dilemma (like how personally involving and emotive the scenario is) either a deliberative or automatic brain process would take precedence over the other (Greene, 2016). Greene (2016) asserts that the automatic and emotive process corresponds with deontological decision making, where rightness or wrongness

are dependent on how the action is initially perceived or characterised, and not the consequences. He bases this assertion on the fact that certain brain areas associated with automatic and emotive processing become active when the participant makes a typical "deontological" type moral judgement in response to a moral dilemma. For example, this would be the case when participants opt not to push the man off the bridge in the overbridge variant of the trolley problem, where pushing the man off would have killed him but saved five others. By contrast, Greene (2016) views decisions that involve deliberative and cognitive processing as reflecting utilitarian reasoning, which involves consideration about the consequences, i.e. consideration of whether the action maximises "utility". He bases this assertion on the fact that when participants make typical utilitarian decisions in response to a moral dilemma, brain areas associated with exercising cognitive control and higher-level cognitive operations become active. For example, this would be the case when participants opt to push the man off the overbridge in order to save five others. Thus, Greene's approach reflects an assumption that common normative ethical theories can be identified with different and dissociable brain circuits. Moreover, he claims that these circuits are inversely related, i.e. when one process is active the other is inhibited/inactive, and that this dynamic corresponds to the different ways ethical dilemmas are often approached (Greene, 2016).

Another way that moral theories can be applied to brain processes is "holistically". So, whereas Greene and others match certain moral theories to particular brain circuits, one might instead view the whole of human moral cognition i.e., all the regions and neural circuits involved, as working together in a manner that corresponds with a particular moral theory. On this approach, it is assumed that the theory that can be shown to "best-fit" with our understanding of these brain processes is more likely to correctly describe human moral cognition, and that this theory is suitable for interpreting subsequent evidence as it emerges. Casebeer and Churchland (2003), as discusses earlier in this chapter, make this claim, and put forward a neo-Aristotelian virtue theory as the candidate for "best fit":

> Our initial take on the domain of moral cognition, then, can and should be informed by a background moral theory (in our case, a neo-Aristotelian virtue theory, according to which moral concerns relate to what we have to *think* and *do* so as to *function well* as human beings).

(Casebeer and Churchland, 2003, p. 172)

Their adoption of virtue theory is partly based on recent neuroscientific evidence, similar to what was discussed in the previous chapter, which appears to indicate that moral cognition is multifaceted, involving "reason, appetite, emotion and affect" (Casebeer and Churchland, 2003, p. 172). They argue that these findings are more consistent with virtue theory than other mainstream normative ethics theories, as virtue theory contends that right action must be understood with reference to the overall functioning of the person, i.e. what must they do to be a "virtuous" person (Casebeer and Churchland, 2003, p. 172). They go further and claim that moral cognition should be informed with a "background moral theory", which they argue should be virtue theory, meaning that any new data should be interpreted in the light of that theory.

There are several problems with this approach that Greene and Casebeer and Churchland have taken up. Firstly, it can be difficult to prove that certain responses participants give to moral dilemmas in fact reflect processing akin to a specific normative ethical theory (Kahane, 2015). For example, suppose I constructed a study that uses the bridge variant of the trolley problem as a moral dilemma, and presented this to participants when they underwent fMRI (this is a similar methodology to the study by Greene at al. (2001)). Suppose I assume that when a participant opts to push an overweight individual off of a bridge to stop the trolley from killing five others that they are making a "utilitarian" type decision, whereas if they opt not to push the individual, they are making a "deontological" type decision. Such an assumption is very common in the literature (Greene et al., 2004; Koenigs et al., 2007; Koenigs et al., 2011; Zheng et al., 2018). This methodology reflects a general consensus in the philosophy literature about which responses are typically supported by certain normative ethics theory, but this does not show that the participants in these experiments are thinking in terms of those theories. The connection between the participants' responses and the corresponding ethical theories may only be superficial. This problem might arise with any theory, though Kahane (2015) singles out utilitarianism as being particularly poorly reflected in the moral judgements made in response to sacrificial dilemmas such as the trolley problem. He points out that utilitarianism is primarily two things: it is maximising, meaning that it aims to achieve the highest value of utility, and it is impartial and unbiased (Kahane, 2015). A participant may give a typical utilitarian response to a sacrificial dilemma, when their

actual thinking may have been motivated by self-interest (many sacrificial dilemmas used in these studies involve danger to the participants as well). One such common dilemma described in Greene et al. (2004) is as follows:

> Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside, you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth, his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others, you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople?
>
> (Greene et al., 2004, p. 390)

This dilemma is subtly different to a trolley dilemma in two important ways. The participant is in danger, and the baby would die anyway regardless of the decision (whereas in a trolley dilemma one individual could die instead of a group of others). So, in this example the participant may be motivated by self-interest, or they may be willing to sacrifice their baby because they were motivated by "group solidarity". Neither of these motivations would demonstrate utilitarian thinking. For many of these sacrificial dilemmas there are a number of possible motivations for coming to a decision that are not typically utilitarian or deontological.

Further to this point, it has been found that psychopaths are more likely to make utilitarian type decisions when faced with sacrificial dilemmas (Koenigs et al., 2011). Psychopaths of course typically lack concern for the well-being of others and are self-centred (and so are not impartial). It is therefore unlikely that they are motivated by the need to maximise happiness according to utilitarian reasoning. Instead it appears that selfishness and callousness (indifference to suffocating the baby in the above example), may be the real motivating factors for their decision.

As noted, Kahane (2015) directs this criticism at utilitarianism. The point may be applied to other applications of moral theories in the interpreting of neuroscientific data. If, for instance, participants give typical deontological responses to a trolley

dilemma (perhaps by opting not to push the overweight person off the footbridge to save five others), we cannot simply assume their decision is attributable to deontological processing. Instead of believing that they have a categorical duty not to kill anyone (a typical deontological reason), it may be that they were motivated by a simple emotional aversion to killing, or that they just did not want to get their "hands dirty" (a far more selfish motivation), or that they simply opted out of the dilemma by doing nothing at all (because the scenario was too distressing). Thus, caution is required when attributing specific kinds of reasoning to participants on the basis of their decision.

In response to the criticisms, Greene might argue that normative ethics theories are manifestations of neural networks in moral cognition. For Greene (2008) the conventional philosophical definitions of utilitarianism and deontology are of secondary importance. What is primary in his analysis are the functional roles of utilitarianism and deontology, i.e. the ways that they generally tend to operate (Greene, 2008). Utilitarianism generally allows for the sacrifice of one individual to achieve a "greater good", and deontology generally condemns the use of any individual as "a means to an end". He argues that it is this general operative functioning that determines whether a decision made by an individual is utilitarian or deontological, as opposed to adherence to philosophical reasoning.

There are several problems with Greene's position. Traditionally, deontology has been associated with reason and reflection as opposed to emotion and reflexivity. For example, Kant's moral system – commonly put forward as a preeminent deontological theory – claims to show why moral obligations must be based solely in reason (Kant, 2002). However, Greene proposes the inverse: that deontological judgements stem from an automatic and emotive brain process, and utilitarian judgements on the other hand, stem from a deliberative and cognitive process. This makes it unclear what is conveyed by Greene's use of theoretical terms. When philosophical theories are employed to describe brain processes, to which they only bear superficial likeness, the link is weak at best.

In reviewing the neuroscientific models of moral cognition earlier in this section, it is obvious there is a pervasive link between them and theories in moral philosophy. Often, it is simplified or traditional versions of these theories that are used,

for example, traditional Kantian deontology and classical utilitarianism, which have clear cut objectives and features. Such theories are used to describe brain processes involved in moral cognition. However, researchers are often reaching when they claim that these traditional theories are manifestations of brain processes; primarily this is a methodological issue, as it is difficult to prove that participants particular responses to ethical dilemmas reflect processing akin to a specific normative ethical theory. Despite this issue, moral philosophy remains important in the interpretation of findings in moral cognition, because it has provided researchers with many diverse and complex frameworks through which to view morality. So, what is needed is a more nuanced use of morally philosophy, that moves away from simplistic and historical versions of theories and towards versions that are neuroscientifically informed.

2.2 Challenge of Circularity in the Neuroscience of Moral Cognition

The neuroscientific study of moral cognition is particularly susceptible to a problem of circularity. In this section I will discuss the general difficulties in conceptualising morality and moral cognition, which are due to the diverse range of morally relevant behaviours. Following this, I will discuss the challenge of circularity which broadly occurs in two directions. The first direction concerns how one's conceptualisation of moral cognition can influence which findings emerge and their interpretation. In order to give study a clear focus, one is required to produce a definition of moral cognition that is unambiguous and methodologically viable. However, this preconception of moral cognition segregates which behaviours are regarded as being morally relevant and hence which brain regions will be focused on. The second direction involves reading a favoured model or theory of moral cognition into the data (which is often flexible and susceptible to the influence of theory). Because of these factors, the study of moral cognition is particularly susceptible to circularity. To overcome this problem the study of moral cognition requires greater reflection and researchers should be more aware of the presuppositions taken.

2.2.1 Difficulties in Conceptualising Moral judgement

The general idea of "morality" is interconnected with ideas of moral judgement and moral cognition. A definition of one entails a definition of the other two. In simple

terms, moral cognition is the operation of the cognitive faculties involved in morality, whether that involves making moral judgements, moral motivations and moral emotions. The borders of morality are notoriously hard to define, partly because they seem dependent on interpersonal and intercultural differences. Typically, studies of moral cognition do not focus on defining morality, but rather rely on presuppositions or a consensus definition. Paradigmatic examples of morally relevant conduct, such as altruistic behaviour, fairness and justice are incorporated into this general definition (Decety and Yoder, 2016; Moll et al., 2008). Without such presuppositions, the study of moral cognition as a branch of scientific inquiry could never get off the ground. Opinions from moral philosophy often influence such presuppositions.

Let us examine an example definition of "morality" commonly used in the neuroscience. Moll et al. (2008) has adopted an operative definition of morality as "the sets of customs and values that are embraced by a cultural group to guide social conduct". An obvious feature of this definition is the inclusion of cultural variability, which is seen as an advantage as it avoids arbitrary dismissals of such variations. This definition also allows for the inclusion of different psychological domains potentially involved in morality, i.e. "disgust, harm, care, fairness and authority" considerations (Moll et al., 2008). Moll et al. (2008) stress that morality is evaluative and requires motivation and presume that it has emerged by way of gene-culture coevolution. This kind of definition has massive scope. Specificity is required to constrain the area of inquiry.

Definitions of "morality" inform definitions of moral cognition and moral judgement. This involves determining which actions, consequences, emotions and character traits are to be regarded as relevant to moral cognition and which are not. For example, a filthy toilet and a brutal murder can both generate strong reactions of disgust in an observer, yet one reaction is morally relevant and the other not. Similarly, what differentiates a judgement about the correct move in a chess game, from the correct action in a trolley dilemma? Questions such as this must be answered before embarking on an empirical study of moral cognition. Similarly, moral judgements can involve a highly diverse array of situations. Studies examining moral cognition have presented participants with moral situations such as: Observing a poor abandoned child in the street (Moll et al., 2002), making choices in varieties of the trolley problem (Greene et al., 2001), observing an individual either intentionally or unintentionally hurting

someone (Decety and Cacioppo, 2012), and judging whether an incestuous relationship between consenting adults is wrong or not (Haidt, 2001). We can regard all of these situations as being morally relevant, yet it seems, superficially at least, that they differ in significant ways. For example, the first situation involves responding to an emotionally troubling image, the second involves weighing up competing emotions and reasons to solve an abstract dilemma, the third involves quickly categorising and judging an agent's actions, and the fourth involves reflecting on a deeply ingrained socio-cultural norm. This indicates the range and complexity of moral cognition as a field of study.

This range and complexity presents anyone conducting neuroscientific research into moral cognition with a dilemma. To construct an experimental methodology a concise and well-circumscribed definition is required. However, in formulating such a definition one risks oversimplification. A possible solution to this problem is to break moral cognition down into many different types of judgements and study each of these types accordingly (Greene, 2015). Much of the literature is already diverse, as the above four very different scenarios presented indicate, and involves a wide range of methodologies. However, there is still a need for an overarching definition of moral cognition that can encompass this diversity, if we regard this research as studying the same thing, viz. "moral cognition". This overarching definition should also guide decisions about what future research should and should not be considered within this category.

If in the future, a practically "complete neuroscience" was achieved, meaning that the brain processes involved in any potential behaviour in differing contexts could be completely explained, one would still need to define what counts as "moral cognition". This conceptualisation process will remain difficult, because of the diversity of behaviours and contexts that are deemed as morally relevant.

Let us take some examples of existing attempts at conceptualising moral judgement in the literature. Alongside their conception of morality, Moll et al. (2008) have a related conception of moral cognition (from which we can extract a definition of moral judgement). They propose that the most distinctive aspect of moral cognition is that it altruistically motivates social behaviour. This definition distinguishes moral cognition from other non-moral forms of social cognition that may have different

motivations, such as selfishness or reciprocal altruism (altruism performed with the expectation of reciprocation). On this definition, whether one's judged permissibility of an incestuous relationship, or one's response to a trolley dilemma are instances of moral judgment depends on whether the judgements involve genuine altruistic considerations. If so, then in studying an individual's responses to these scenarios one is studying "moral cognition". On this definition morally relevant social conduct can be distinguished from non-moral social conduct by examining the motivations of the conduct.

Greene (2014) uses a different definition of moral cognition. He claims that "morality is a set of psychological adaptations that allow otherwise selfish individuals to reap the benefits of co-operation". In explaining this, Greene (2014) describes morality as an evolutionary solution to the "tragedy of the commons". The tragedy of the commons is a societal problem where an individual acting in their own selfish interest undermines some common, social good, and thereby undermines her own interests. In other words, when everyone acts selfishly, the common good – the "commons" – become spoiled. Compared to Moll et al.'s definition of moral cognition, Greene's is far broader in the sorts of behaviours it implicates. For example, motivation is still important, as genuine altruism may be a mechanism to combat selfishness and avert the tragedy of the commons. However, it is only a means to the end that is human cooperation. This definition can help sort morally relevant social conduct from non-moral social conduct by examining the general functioning of the given behaviour at an individual and a societal level.

As a brief aside, a significant omission in Moll and Greene's definitions is the normativity inherent in moral judgements. Normativity concerns the sense of "ought" behind moral judgements, and the idea that there is a justification for the claims that morality makes on us (Korsgaard, 1996, pp. 9-10). Neuroscientific studies and articles like those of Moll and Greene seek to provide an explanatory account of morality, i.e. a causal description of how moral judgements are made and why individuals are susceptible to the influences of morality. In doing so they regard the question of whether a person actually "ought" to act morally as either secondary or irrelevant. I will discuss this distinction further in the next chapter.

2.2.2 Circularity in Conceptualising Moral Cognition and Interpreting Data

When applied to experimental design both definitions mentioned in the previous section bring us to a fundamental methodological challenge: The problem of circularity. The way moral cognition is framed explicitly and implicitly influences which cognitive faculties will be examined and how any findings are interpreted, and hence which will constitute an important part of an emerging model of moral cognition.

For example, because Moll et al. (2008) focus on altruistic motivations, any neuroscientific findings will be interpreted in this context. In this case, moral cognition becomes the study of the brain processes that constitute such motivations. Moll et al. (2008)claim that the literature suggests that emotional and cognitive faculties in moral cognition are non-dissociable. In their view, moral motivations are constituted by the integration of these processes, which form cognitive-emotional association complexes, represented in cortico-limbic neural assemblies (Moll et al., 2008). A conflicting moral decision represents a conflict between two different motivations, each represented by different cortico-limbic assemblies. The researchers particular focus on motivations and their interpretation of the data trace back to their initial conception of morality.

In contrast, as Greene's definition of morality centres on the psychological adaptations that enhance co-operation, he views the evolution of these psychological mechanisms as important to his theory (Greene, 2014). Greene proposes the "the dual-process model of moral judgement", whereby we have a fast, automatic and largely emotive way of making moral judgements, and a slower, deliberative and more energy intensive way. Greene (2014) pitches this dichotomous model for moral judgement as a classic evolutionary trade-off between efficiency and flexibility. It makes sense from an evolutionary standpoint that we would have evolved an automatic way of making moral judgements for maximum efficiency, as well as a more deliberative and mentally costly way for harder and novel moral dilemmas. The key point here is that Greene's interest in such psychological adaptations, and his dual process theory, trace back to his initial conception of morality.

In discussing the inherent circularity of research into moral cognition, Casebeer and Churchland (2003) claim that the theory of morality used to create a working definition of moral judgement dictates "how widely you cast your moral cognitive net".

For example, if one defines moral cognition in a given way, based off a theory of morality, then this will produce some initial empirical findings. These findings can aid in the development of models of moral cognition and even eliminate some models that are unsupported and deemed unrealistic. This process of reformulation can influence changes the initial working definition of moral cognition. However, Casebeer and Churchland (2003) also claim that this circle is not necessarily vicious, and instead propose that definitions of moral cognition and experimental models need to be developed together. Hence, this process does have a circular element, which is inherently problematic. However, if this analysis becomes more reflective, it may be reviewed as hermeneutical. With a nuanced use of moral philosophy and awareness of one's presuppositions, this circularity will be far harder to miss.

In the previous chapter I showed why it is widely accepted that there is no specific module in the brain that is solely responsible for moral cognition. Instead what exists is an integrated network of brain regions, that can be implicated in a wide variety of studies, that employ differing methodologies that examine different kinds of moral judgements. Greene (2015) concurs with this point and argues that moral cognition is "fragmented at the cognitive level but unified at the functional level". However, whatever "function" is used in the working definition of moral cognition will dictate which fragmented cognitive components will be tested and focused on.

So far, what has been discussed is how initial conceptualisations of moral cognition can influence the methodologies chosen, which brain regions are studies and how any findings are interpreted. Another aspect of the challenge of circularity is the tendency for researchers to read their favoured model of moral cognition into the neuroscientific data.

One reason for this circular tendency is that much of the neuroscientific evidence is ambiguous in relation to these models. In other words, it is often the case that evidence can be used to support most if not any of them. In section of "Moral Brains" (2016), the philosopher Jesse Prinz performs a brief literature review of the neural correlates of moral judgements and sets out the various models of moral cognition. Before launching into their descriptions however, he states:

> To put it bluntly, every model presented here is consistent with every study cited in the previous subsection.

(Prinz, 2016)

Prinz argues that we cannot be decisive in any endorsement of a model of moral cognition based on the neuroscientific literature alone. Nevertheless, he presents his own sentimentalist model as having the greatest support from the literature. Prinz (2016)does so by presenting an alternative method of analysis, where theories of moral cognition, often originating from moral philosophy are utilised to interpret findings. Though this approach may be useful, it should be used with caution. The fact that neuroscientific studies can be interpreted flexibly with regards to moral theories, means that it is very easy to read a moral theory into the data. Prior theoretical positions therefore would be more likely to dictate one's conclusions than the findings themselves.

Let us take the example of Prinz's own analysis of one study and its relation to models of moral cognition. We will see that Prinz's analysis is as problematic as the others he criticises. The study he addresses concerns a neuroimaging experiment by Decety and Cacioppo (2012), which was discussed briefly in chapter one. In this study, high-density Event Related Potentials (ERPs) were used, which represent participants' brain activity in response to a fixed event, such as when participants watch a video. Participants were made to watch a video depicting an actor either intentionally or intentionally harming another individual. Intentional harm ERPs were distinguishable from unintentional harm ERPs at three temporal points, which correspond to activity in the right pSTS first, the amygdala second, and finally the vmPFC third.[10] The researcher's interpretation of their own findings is that early in moral cognition processing, the intentions of the actor are evaluated, and this guides the judgement. Following this is the involvement of affective processes, which act as a "gain antecedent", alerting the observer to the "moral salience" of the scenario, before a harm evaluation is made (Decety and Cacioppo, 2012).

Prinz (2016) disagrees with this interpretation as it relegates the role of emotions in moral judgement, which are proceeded and guided by a seemingly cognitive process, i.e. the intentionality judgement. Using the same data, Prinz argues that any model of moral cognition could account for the initial involvement of this

[10] In chapter one, all three of these areas were discussed, and are implicated in evaluating intentions, emotional arousal and decision-making respectively.

cognitive function. He proposes a sentimentalist "constitution" model, which asserts that sentiments (dispositions for certain emotions) constitute moral judgements (Prinz, 2016). For example, if one sees someone intentionally harming another, as in this experiment, one gains a sentiment which disposes one to anger, which is equivalent to a moral judgement that the action was wrong. In Prinz's model, a categorisation of the act by evaluating the intentions of the actor is necessary before one can have an emotional reaction to it, and hence form a judgement. Thus, the early involvement of a cognitive brain region, he argues, is no real threat to his model. In Prinz's analysis, the involvement of affective brain regions represents more than just a "gain antecedent", instead they are the judgement itself.

There is a third way to interpret these results, which rejects a hard dichotomy between reason and emotion, and argues for a neuroscientifically integrative view. This is the position taken up by this thesis and will be discussed later in this chapter. In this view, brain regions implicated in moral cognition function as a network, so it does not make sense to say that any one region of kind of process solely constitutes the moral judgement. Instead, occurring very early in the moral judgement is the integration of affective and cognitive processes to carry out a judgement. Decety and Cacioppo (2012) hint towards a seemingly integrative process when they emphasise the reciprocal connections between the pSTS and the amygdala, and the projections of the pSTS and amygdala to the vmPFC.

Prinz (2016) argues that the ambiguity of the available empirical evidence such as in the Decety and Cacioppo (2012) study indicates the need for interdisciplinary work. He argues that the interpretation of findings in empirical neuroscience needs to be supplemented by "philosophically grounded theories" and "behavioural work", which is sometimes but not regularly the case in current neuroscientific studies (Prinz, 2016). I concur with Prinz's conclusion, however as an amendment I want to empathise the importance of theoretical neuroscientific work, in this interdisciplinary process. Theoretical neuroscience, for example the work in the article by Gillett and Franz (2014), is few and far between, and ought to be a primary focus in the field of moral cognition. It also requires emphasis that this interdisciplinary work should be mutual. In other words, the collaboration between moral philosophy and theoretical neuroscience should not exist in a unitary direction.

The study of moral cognition is particularly susceptible to circularity. This is due to the need for presuppositions in conceptualising moral cognition, and in the use of theory to interpret of any findings, which are often ambiguous and open to multiple interpretations. To overcome these problems, a reflective collaboration between moral philosophy and theoretical neuroscience is required to inform the neuroscience. Most importantly, this must be accompanied by an awareness of the origin of any philosophical or theoretical presuppositions and how these may influence findings.

2.3 Methodological Limits in the Neuroscience of Moral Cognition

In chapter one, I examined key studies in the neuroscience of moral cognition and discussed their implications for the nature of the moral brain. The moral network is highly disseminated in the brain, involving a multitude of brain regions with various functions. This presents a problem for the cognitive neuroscience of morality, as most neuroimaging techniques are limited in one or more ways in their ability to capture the moral network. fMRI (functional Magnetic Resonance Imaging) and EEG (electroencephalography) are the two most common neuroimaging techniques used to study moral cognition, though TMS (transcranial magnetic stimulation) is also common. fMRI has high spatial resolution, and poorer temporal resolution, this means that it can distinguish brain activity in specific brain regions more clearly. However, it provides less data on the precise timing behind that activity. The inverse is the case for EEG which has low spatial resolution and high temporal resolution (Mulert et al., 2004).

When the data from each of these techniques are analysed together it can help elucidate both the timing and the localisation of brain activity. However, due to the nature of moral cognition, this will continue to yield an incomplete picture. If, as the literature suggests, moral cognition involves a network, then it must be understood somehow as integrative. Hence, current neuroimaging techniques, that focus on specific brain regions will be of limited value. There are some techniques that can be applied to neuroimaging data during the analysis phase that provide some information on the interaction of multiple brain regions, however this remains limited. Currently, the only way to address this technical limitation is through theory.

In this section I will describe these inadequacies in more detail, which can be explained in three parts. Firstly, because of the context-dependent nature of moral cognition, many neuroscientific experiments will lack ecologically validity, meaning that they fail to emulate the real-life conditions that moral judgements occur in. Secondly, methodologies and experimental techniques fail to capture the temporal nature of moral decision-making, that can occur over minutes, hours and days. Thirdly, the same methodologies and experimental techniques will fail to capture the holistic nature of moral cognition, that operates as a complicated network of processes.

2.3.1 Ecological Validity in Testing Moral Cognition

A common criticism of experiments employed in the neuroscience of moral cognition is that their methodologies are not ecologically valid. Ecological validity refers to how well a methodology resembles the real-life conditions the experimental target occurs in. In experiments on moral cognition, the sorts of moral judgements a participant makes in experimental conditions may differ from moral judgements made in everyday circumstances.

One important example of a lack of ecological validity is how experiments often fail to emulate the context-sensitivity of moral judgements. A rigorous and replicable experimental design carried out in a laboratory environment can lack the deep contextual setting most real-world moral judgements are made in. This contextual richness might involve a range of factors. Moral judgements can be intensely emotional, especially if they are personally engaging and the real-life stakes are high. Also, moral judgements are embedded in a social context which involves interacting with others who may be connected by complex relationships. Finally, the physical and cultural environment a moral judgement is made in can radically alter our ethical conclusions, as the same action performed in two different contexts can lead to either condemnation or endorsement. For example, consider the actions of Hannie Schaft and the Oversteegen sisters, who were part of the Dutch resistance to Nazi occupation during World War II. They seduced and killed German soldiers and sabotaged various military targets. In the context of Nazi occupied Netherlands these three women are war heroes and their actions praised. In a different context their actions would be condemned. Casebeer and Churchland (2003) identify a range of ways that moral cognition might be context dependent, and discuss the implications this has for

empirical research into moral cognition. These include the inability to study certain kinds of genuine moral behaviours, such as moral heroism or akrasia (weakness of the will). They also point out that our moral judgements are organic and directed, meaning that in addition to being context sensitive, they involve a sense of having to act. A judgement about how one will *actually* act may be neuroscientifically very different from a judgement made from the experimental "armchair" (so to speak) (Casebeer and Churchland, 2003).

In general, methodologies that contain more abstract moral dilemmas with significant imaginary components are more susceptible to problems of ecological validity. This is particularly true of experiments that utilise sacrificial moral dilemmas. In such experiment's participants are asked to make a difficult moral choice, often represented by sacrificing or harming one individual to achieve a greater good (Greene et al., 2001; Greene et al., 2004; Koenigs et al., 2011). The most typical example of these dilemmas are trolley problems, including both the switch and bridge variants.[11] These experimental designs are abstract, highly hypothetical and unlike the kinds of moral dilemmas people encounter in real life. While such dilemmas may have a place in the overall analysis of moral cognition, some claim the backbone of moral cognition research should instead focus on more realistic everyday examples of moral judgement (Kahane, 2015).

In response to the problems outlined here, a number of abstract and implausible sacrificial dilemmas have been dropped from experiments in favour of more ecologically valid alternatives. For instance, Feldmanhall et al. (2014) only utilised 15 moral scenarios from previous literature and developed 50 more scenarios of their own to address this issue. They argued that many moral scenarios in the existing literature were too extreme and unfamiliar in nature, giving the example of "deciding whether to cut off a child's arm to negotiate with a terrorist" (FeldmanHall et al., 2014). Their aim was to include moral dilemmas that had higher ecological validity, i.e. scenarios that were more familiar and relevant to the participants understanding of ethical rules. This move towards greater ecological validity is an important step in increasing the quality of research in the neuroscience of moral judgement.

[11] Refer to chapter one for descriptions of the trolley problem and its variants.

From an experimental perspective, the lack of ecological validity i.e. a reductive approach, may have certain advantages. In a criticism of the methodologies employed in the neuroscience of moral cognition, Kahane (2015) points out that an abstract and artificial moral dilemma might allow for the isolation of the specifically moral components in cognitive processing. In other words, by filtering out factors that would be present in all kinds of context-dependent social interactions, it might be possible to identify certain moral responses that do not rely on mere "social convention". There may well be a tension between ecological validity and attempting to isolate certain brain processes such as this.

However, these potential downsides of ecological validity can be addressed in other ways, that do not include sacrificing it altogether. For instance, including a broad array of methodologies ensures that the cognitive complexity of moral judgements in many different contexts can be assessed, and this includes the kinds of judgements that have higher or lower reliance on social convention. For example, studies that examine automatic and social judgements can be analysed in conjunction with studies that examine difficult moral judgements, where the reliance on "convention" would be less.

2.3.2 Challenge of Temporality

In order to study moral cognition, there is a need for a consistent and replicable behaviour that can be repeated and recorded accurately. Hence, to study the neural correlates of a moral judgement, what exactly is considered as a moral judgement needs to be standardised and simplified. For example, in chapter one, many studies examined participant's judgements in response to sacrificial dilemmas in a standardised way. In one such study by Greene et al. (2004) participants were given 46 seconds to read the scenario (depicting the moral dilemma) and give their response. Eight fMRI images were taken around the time of the response (four prior, one during and three following), capturing a 16 second window around the judgement. fMRI measures differences in haemodynamic flow (blood flow), which is (under a well-accepted assumption) related to brain activity. Greene et al. (2004) explains that the images taken following the response were taken because they allow for a "lag in h[a]emodynamic response".

Methodologies such as this are common, and only capture the brain during the few seconds of making a moral judgement. In many cases this limitation is unproblematic, as the target brain activity corresponding to a behaviour or cognition is

highly specified and well defined. However, moral cognition is a complex process. In everyday life moral judgements can be formed over the course of minutes, hours and even days. Operating morally can involve interacting with others socially and involve reflecting on moral problems in relation to past and current experiences. Thus, moral cognition cannot be adequately captured in the few seconds of brain activity recorded during neuroimaging studies, which remains a methodological necessity. Furthermore, regarding ecological validity, moral judgements are formed within the broader context of one's moral life. However, during neuroscientific experiments, participants are often asked to read a scenario, merely imagine it unfolding, and respond in a standardised way.

Pizarro and Bloom (2003) discuss the temporal range of moral cognition in relation to reason. They do so in a response article to the work of Jonathan Haidt, who claims that reason is only of secondary importance in moral judgements (Haidt, 2001). Pizarro and Bloom (2003) argue that reason is extremely important, and that its role is just less visible. What is highlighted is how morality cannot be reduced to the brain activity or behaviours surrounding a specific judgment. They argue that cognitive appraisal, the cognitive interpretation of a scenario, can greatly influences one's intuitive judgements. For example, searching for more contextual information about a seemingly rude individual, can reveal that they had just lost their job, which alters your perception of them. Secondly, through a selective control over one's environment, controlling what one is exposed to and what one learns, can '"educate' the moral intuitions" (Pizarro and Bloom, 2003). An example given by Bloom and Pizarro (2003), is how one can consciously choose to take a class on racism, taught by an African American professor, which involves a deliberate and conscious decision to change one's intuitions. In these ways, prior reasoning can inform one's automatic judgements, but in a subtle way that is unobservable in neuroscience.

Think of the topic abortion. It is true than many people have intuitions about abortion that guide their judgements in most circumstances when the topic comes up. Individuals even commonly identify as being "pro-life" or "pro-choice". If one wanted to examine the neural correlates underlying moral judgements about abortion then, one may easily conclude that emotive and automatic processes contribute to these judgements. In the few seconds surrounding the judgement, neuroimaging may reveal that this is the case. However, one's views on abortion are not reducible to a judgement

in that instance. Instead, if I were to interview participants about their views on abortion, they would reveal the intermittent prior reasoning lead them towards their intuition, which may have occurred over the course of months or even years. Perhaps they have even had personal experiences with abortion that have led them to their stance.

I call this inability for neuroscience to capture the temporal nature of moral cognition, the challenge of temporality. However, it is important to remember that although the current neuroscience literature is inadequate in this way, the results remain important in contributing to our understanding. Such findings give us a general, albeit incomplete idea of the brain processes involved in moral cognition, in particular during "short-term" moral judgements. To describe the expansive temporal nature of moral cognition an interdisciplinary analysis is required, that involves behavioural work and theoretical neuroscience. Experiments in social psychology are helpful as they allow researchers to study moral judgements that extend further temporally, although they provide limited information about specific brain regions involved.

2.3.3 Challenge of Holism

As outlined in the previous chapter, much of the empirical work studying moral cognition has focused on particular brain regions and their implicated functions. This is necessary work. But if – as the evidence seems to indicate – these brain regions interact in complex ways as a network, this work only reveals half the picture. Moral cognition is a complex process that involves diverse brain regions, disparate in space and function, that build off their domain general functions to operate as a network. There is a general and well-accepted assumption is that there is no brain module that is solely responsible for moral cognition (Gillett and Franz, 2014; Greene, 2015). This indicates the need for a shift in approach, towards neuroscientific research that examines how the implicated brain regions interact. Despite this assumption, researchers continue to search for a specific neural correlate of moral judgement. For example, after discussing various models of moral cognition, Prinz (2016) raises a question for these alternatives compared to his "constitution model", that was discussed earlier in the chapter:

> If moral judgements are not emotional states, what brain structure is their neural correlate? There is no obvious candidate suggested in the literature. We are left wondering where moral judgements reside, with no clear

> proposal how to find the answer. The constitution model provides an answer: moral judgements reside in emotion pathways, or, more accurately, in the joint activation of those pathways and brain structures that represent actions.
>
> (Prinz, 2016)

This is the type of view that needs to be countered. Why do moral judgements have to be reduced to a simple neural correlate? This reduction of morality may seem to provide a simple explanation, but it misses a great deal. Instead, moral cognition in the brain operates as a network, that cannot be reduced to a single circuit. Many brain regions and functions are implicated in moral cognition, and are all the "neural correlates", thus, it makes no sense to single out a single process in order to explain what a moral judgement is.

Hence, to better describe the relationship between the brain and morality, the study of moral cognition needs an increased emphasis on taking a neuroscientifically integrative perspective. This means that the focus shifts from studying the individual functioning of the various brain regions which comprise the moral network, towards an account of how they communicate to carry out their collaborative higher-level function of moral cognition. Many basic foundations of a neuroscientifically integrative moral cognition were referenced from the literature in chapter one. These included that moral cognition involves diverse and disparate brain regions, in both space and function, that build off their domain general functions and integrate at multiple levels. Furthermore, there has been no specific "moral module" discovered in the brain, which has led Greene (2015) to predict that:

> …moral cognition will continue to flourish, not as the study of a single cognitive organ (Hauser, 2006), and not only as the study of loosely related problems, but as a testing ground for more general questions about the nature of high-level cognition, questions about how the brain's disparate cognitive components are integrated to produce (mal)adaptive behavior.
>
> (Greene, 2015)

Taking an example of a higher-level theory of cognitive function, Gillett and Franz (2014) have proposed using the work of John Hughlings-Jackson as a framework to examine moral cognition in a more neuroscientifically integrative context. Hughlings-

Jackson was an English Neurologist prominent in the late 19th century, who proposed important ideas about complex brain organisation (Franz and Gillett, 2011). His fundamental idea was that higher-brain areas elaborate the functioning of simpler brain areas. In other words, through successive levels of integration, simple and specific sensorimotor brain process are represented and then re-represented by different and more complex sets of brain regions, which converge on the brain's highest centres.

Franz and Gillett (2011) have used Hughlings-Jackson's thesis of brain organisation to analyse seemingly contradictory results from recent cognitive neuroscience studies that examined social cognition. In such studies many well-circumscribed and specific brain regions (such as the TPJ) have been implicated in a diverse array of higher functions. At first it seems paradoxical that a specific brain region would contribute to each of these diverse functions, however Franz and Gillett argue that in the light of Huhglings-Jackson's ideas these results are more explicable.

For example, in chapter one I introduced the meta-analysis by Decety and Lamm (2007), which analysed 70 functional neuroimaging studies of the right TPJ. They found that the right TPJ was implicated in theory of mind, empathy, attribution of agency and attentional shifting. In reference to the TPJ, Gillett and Franz (2011) state that basic sensorimotor processes are being integrated and re-represented in this brain region, by noting that each of these social functions require the integration of contextual information so the actor can respond to complex situations. Decety and Lamm (2007) propose an idea very similar to that of Hughlings-Jackson when interpreting their own results:

> …activation in the TPJ during social cognition may therefore rely on a lower-level computational mechanism involved in generating, testing, and correcting internal predictions about external sensory events. Such an interpretation is consistent with an evolutionary view that higher levels operate on previous levels of organization and should not be seen as independent of, or conflicting with, one another. Evolution has constructed layers of increasing complexity, from nonrepresentational to representational and meta-representational mechanisms, which need to be taken into account for a full understanding of human social cognition.
>
> (Decety and Lamm, 2007)

Gillett and Franz (2014) have also applied the ideas of Hughlings-Jackson theoretically to moral cognition more specifically. They proposed that from a high-level cognitive perspective, moral reasoning would involve the maximal integration of emotion and social cognitive structures, based on the representation and re-representation of environmental contingencies. Meaning that moral cognition is a complex and essentially integrative process. Hughlings-Jackson's ideas are important for an "integrative moral cognition" because they set out a framework that shifts focus from individual brain regions and processes towards an account of the "moral network" as a whole.

Viewing the brain at this level naturally undermines the dichotomy between reason and emotion. Specific brain regions, like the amygdala may primarily function as emotional responders, and others as purely cognitive structures, however, this distinction breaks down when examining each region as part of an integrated network. Instead emotion, feeling and reason are intertwined, and are essentially related during the higher processing of moral cognition.

I call this the challenge of Holism. Empirical neuroscience alone cannot adequately describe the holistic nature of moral cognition. In the current neuroscientific literature there needs to be an emphasis on testing theories of higher-level cognition. This work would greatly benefit from the accompaniment of theoretical work in both neuroscience and philosophy. Importantly however, the technological limitations preventing neuroscience from describing the holistic and temporal nature of moral cognition may one day be overcome. It may be possible, in the near future to accurately study higher-level cognitive operations in the brain and the complex relationships between regions and processes. Then empirical neuroscience may play a greater role in describing the moral network as a whole.

2.4 A Call for an Interdisciplinary and Integrative Approach to Moral Cognition

In this chapter I have discussed three groups of challenges facing the neuroscience of moral cognition. Interspersed throughout the chapter were also criticisms of some predominant models of moral cognition. The first group of challenges included issues with the use of moral philosophical theories in the neuroscience of moral cognition.

Models of moral cognition have roots in such theories, however their use of moral philosophy is often simplistic and cursory. The second group of challenges concerned issues of circularity in the neuroscience of moral cognition. These arose regarding the conceptualising of moral cognition and moral judgement and in the interpretation of any findings. The third set of challenges concerned the methodological limits of the neuroscience of moral cognition, such as the lack of ecological validity in experiments, and the inadequacy of neuroscience, on its own, to fully explain the temporal and holistic extent of moral cognition.

These challenges highlight the need for an integrative and interdisciplinary approach to moral cognition. By integrative, I refer to a neuroscientifically integrative approach, where moral cognition should be viewed as a complex network of brain regions, that operate together. From this perspective, there is increased emphasis on theoretical work attempting to explore the nature of higher-level brain functioning.

By interdisciplinary, I refer to a synthesis of methods and perspectives to study moral cognition. In this chapter I have shown the dependence of neuroscience on theories in moral philosophy. In order to conceptualise moral cognition, philosophical presuppositions are necessary, and moral theories provide frameworks to aid in the analysis of ambiguous experimental data. However, much of the use of moral philosophy is limited, as there is an excessive focus on simplistic and traditional theories. Instead, neuroscientifically informed variations of such theories should be used. The neuroscience of moral cognition is also particularly susceptible to circularity, in part because of the reliance on philosophical presuppositions. Without an understanding of philosophical theories, it is more difficult to see and reflect on this circularity. Finally, because of the current methodological and technological restraints of experimental neuroscience, theoretical neuroscience, informed by moral philosophy is needed to explore the holistic and temporal nature of moral cognition. In each of these cases a synthesis of perspectives is required. Therefore, an interdisciplinary approach to moral cognition is needed, that primarily uses moral philosophy and theoretical neuroscience alongside empirical neuroscience. Behavioural work, such as in social psychology will also prove useful, as it allows one to study morally relevant behaviours in real-life contexts experimentally. Compared to cognitive neuroscience, social psychology also allows one to study moral judgements made across greater periods of time.

In the introduction I laid out a basic description of what an integrative account of moral cognition might include. Both features above are essential features of this framework. In chapter one I discussed the contributions of neuroscience to our understanding of the relationship between the brain and morality. In this chapter I have discussed where neuroscience can go wrong. In chapter three I will introduce moral philosophy into the discussion. In particular, the focus will be on the work of Christine Korsgaard and her neo-Kantian account of moral obligation. This discussion will reveal that in order to produce a complete understanding of the relationship between the brain and our moral lives, an account of the normativity of moral claims is necessary. Importantly, one cannot give a normative account of morality solely using neuroscience. This is the third and final feature of an integrative account of moral cognition. In chapter four I will conduct an exercise of this interdisciplinary work. This will involve integrating what has been learnt from empirical and theoretical neuroscience and moral philosophy, in order to establish a basic integrative account of moral cognition.

3 Korsgaard and the Normative Question in an Integrative Account of Moral Cognition

The aim of this thesis is to discuss the need for an account of moral cognition that is integrative in character and interdisciplinary in approach. In chapter one, through a selective survey of the literature I showed that moral cognition is a complex process involving a network of brain regions disparate in space and function. These brain regions build off their domain general functions to carry out moral cognition, which involves reason, emotion and social cognition. These findings support a multi-faceted and integrative model. In chapter two, I outlined three challenges facing the neuroscientific study of moral cognition. These included challenges of circularity in conceptualising moral cognition and in interpreting any findings, maintaining ecological validity during experiments, and the inability of current neuroscientific techniques to adequately explore the holistic and temporal extent of human morality. These inadequacies show the need for an interdisciplinary and integrative account of moral cognition.

In this chapter I will introduce moral philosophy into the current discussion as part of this interdisciplinary approach. I will set out a way of understanding moral decision-making that is based in Christine Korsgaard's neo-Kantian moral theory.[12] This account will reveal many important factors that are obscured in the contemporary debate in neuroscience, including a greater emphasis on the role of reason and reflection in moral cognition. Korsgaard's theory of moral decision-making emphasises agential unity, which is consistent with the integrative nature of moral cognition. When choosing how to act Korsgaard claims that one reflects and endorses an inclination as a reason to act. This process is necessarily unifying, as any underlying processing must culminate in a single decision. If, as the literature suggests, the brain regions and processes involved in moral judgement operate as a network, then in some sense they are unified.

Finally, Korsgaard's moral theory provides the appropriate setting to discuss the distinction between explanatory and normative accounts of morality. The goal of

[12] This work will primarily concern Korsgaard, C. (1996) *The sources of normativity.* Cambridge University Press.

neuroscience in moral cognition is to produce a causal description of the brain processes underlying morality. However, there is another important aspect of morality, that of "normativity". Normativity here concerns the sense of "ought" inherent to moral claims on us. Neuroscience alone can never account for normativity, which can only be answered from a first-person perspective. Hence, moral philosophy is needed in a complete and integrative account of moral cognition. In chapter four, I will develop an account of moral cognition by relating Korsgaard's ideas about reason and reflection with modern empirical and theoretical neuroscience.

3.1 Korsgaard's Account of Moral Obligation

In this section I will summarise Korsgaard's account of moral obligation, drawing primarily from her 1996 book "The Sources of Normativity". In this book, Korsgaard sets out to answer what she calls "the normative question". This question concerns the foundations of human morality, and as such goes deeper than much of the scientific work in moral psychology, which merely aims to explain why humans have specific moral practices and beliefs and perform certain actions under the idea of morality. The normative question instead asks: "what justifies the claims morality makes on us[?]" (Korsgaard, 1996, pp. 9-10). In other words, it goes beyond a merely descriptive account of moral behaviours and attempts to explain the normative dimension inherent in moral judgements, i.e. the idea that there are certain ways that we should or should not act. One can attempt to explain the psychological origin of an obligation, such as "I ought to give to charity", but no matter what the explanation, one can always question the legitimacy of this obligation. Hence, the essential question for a philosopher like Korsgaard is, what justifies the authority that morality holds over us in this instance? What is the source of normativity?

Korsgaard (1996) examines historical examples of individual philosophers and their schools of thought that have attempted to answer this question, and then develops her own solution, based primarily in the philosophy of Kant. She presents this solution as a synthesis of the previous historical attempts. In short, Korsgaard argues that normativity arises when an individual reflects upon their thoughts and endorses a reason to carry out a particular action that fits with what she calls their "practical identity". Such acts are autonomous, as by doing so one is "self-legislating", i.e. setting

the rules according to which one will live. I will explain Koragaard's solution in more detail later in this chapter.

In the forthcoming exegesis there will be a significant focus on the moral psychology that underlies Korsgaard's theory. This focus is important as it looks at the intersection between a normative account of morality and work in psychology and neuroscience. These connections will be developed in chapter four where Korsgaard's ideas about rational reflection will be integrated with modern neuroscientific work.

Before embarking on this analysis several clarifications of the overarching aim are required. The following is not intended to be a comprehensive or exclusive model of moral cognition. Several important questions will be left unresolved, and there may be other theories more suitable for addressing these. There are two primary reasons for focusing on this theory. The first is that it reveals a critical link between reason and reflection, that has been largely ignored in the contemporary neuroscientific literature and yet is highly suitable for neuroscientific study. Secondly, it indicates neurosciences inability to describe the normative aspects of morality.

3.1.1 The Normative Question

Ethical claims are normative. This means that they make claims about how we should act, and how we should view the actions of others (Korsgaard, 1996, p. 8). As Korsgaard puts it, they are a kind of "command", though not the kind of command that is followed because of a threat, but rather because they are authoritative. Ethical concepts do not merely describe, suggest or predict how we might conduct ourselves through the world, but rather serve as the basis of our behaviours. Indeed, they seem to carry a unique power, in that their pull to act a certain way can override all other interests. Moral behaviour often involves personal costs, and history provides us with many examples of individuals enduring great hardship or pain, and even sacrificing their lives for moral causes.

One such example are conscientious objectors. Archibald Baxter, the father of the famous New Zealand poet, James K. Baxter, was a conscientious objector during World War One. He was conscripted and subsequently court marshalled for refusing to fight. Following this, Baxter was deported to the Western Front, and due to his continued refusal to take up arms he was beaten repeatedly and tied up close to the enemy lines under heavy artillery fire. Baxter survived unscathed, yet his story remains

a testament to the power morality has over us. Sometimes our moral obligations can be difficult, sometimes even demanding our own lives, yet the obligation stands. The normative question posed by Korsgaard asks where does the authority of such obligations come from?

Korsgaard describes three conditions that must be met to answer the normative question. She maintains that if a solution fails to meet any of these three conditions, it should be treated with scepticism. Firstly, a solution must not be suitable merely from a third person perspective, as the normative question is asked from a personal standpoint. In other words, it concerns one's own obligations. As Korsgaard writes, it should "…satisf[y] us when we ourselves ask the normative question" (Korsgaard, 1996, p. 17).

Secondly, the solution must be transparent. Knowing the evolutionary or psychological origins of our moral intuitions should not disrupt the claims morality makes on us. A good solution should be able to survive the knowledge of where our moral beliefs and sentiments stem from. As Korsgaard puts it:

> A normative moral theory must be one that allows us to act in the full light of knowledge of what morality is and why we are susceptible to its influences, and at the same time believe that our actions are justified and make sense.
>
> (Korsgaard, 1996, p. 17)

Finally, the solution should concern itself with our personal identities. As noted, morality can demand great things from us. For example, Archibald Baxter was even willing to die because of his moral opposition to war. To account for such actions Korsgaard believes that morality must relate somehow to who we are, i.e. to our identity. Hence, a solution to the normative question:

> …must show that sometimes doing the wrong thing is as bad or worse than death. And for most human beings on most occasions, the only thing that could be as bad or worse than death is something that amounts to death – not being ourselves anymore.
>
> (Korsgaard, 1996, pp. 17-18)

3.1.2 Korsgaard's solution

To answer the normative question a theory must address us from a first-person perspective, be transparent, and appeal deeply to our sense of identity. Korsgaard (1996) provides a theory that she believes satisfies these three criteria. In introducing her theory, she reviews other historical solutions to the normative question. I will briefly summarise this review before setting out her theory, as it further clarifies the nature of the problem, and in certain ways mirrors the modern debate.

Korsgaard discusses several attempts at solving the normative question that have arisen in response to what she calls the Modern Scientific World View (Korsgaard, 1996, p. 18). This is the basic view that the world is made of matter and is devoid of any larger inherent purpose. The conventional metaphysics of previous eras, whether that be of Ancient Greece or Christian Medieval Europe, appear more conducive to a form of ethical "realism", i.e. the idea that normative ethical claims are based in something real. For example, one might see a justification for the claims morality makes on people in the authority of God.[13] However, this kind of solution is no longer available to many people following the rise of the Modern Scientific World View in the seventeenth century.

The Modern Scientific World View is an appropriate perspective to take for this thesis, given the central importance of neuroscience, which aims to explain neurological and psychological phenomena in naturalistic, rather than super-naturalistic, terms. If one takes this perspective, it is important for any theory of moral concepts to meet a certain criterion of explanatory adequacy.[14]

Korsgaard presents her selected historical solutions chronologically. Each solution, heralded by a different retinue of philosophers, were all formulated in response to the Modern Scientific World View which disrupted many of the previously conventional "sources of normativity" for moral concepts (e.g. the authority of God). Furthermore, she argues that each subsequent solution was formulated in response to

[13] While this solution may appear more conducive to ethical realism, this does not mean that it is without its problems. The Euthyphro dilemma for example asks: Is the good so, merely because God wills it, or does God endorse the good, because it is already good? If the former is true then the good is arbitrary because God could will anything to be good, if the latter is true then God has no authority over the good, and the question of 'what is good' has not been answered.

[14] Korsgaard calls these the "practical and psychological effects of moral ideas". Korsgaard, C. (1996) *The sources of normativity.* Cambridge University Press. p. 12.

the previous, each attempting to remedy many of the problems faced by the predecessor. I will briefly summarise each of these solutions to illustrate this historical development, before moving to Korsgaard's own solution.

The first solution presented by Korsgaard is "voluntarism", which was a theory first proposed by Hobbes and Pufendorf in the 17th century (Korsgaard, 1996, p. 21). Hobbes and Pufendorf noted that the claims moral concepts place on us are not mere suggestions, instead they more closely resemble commands or laws. Law, they argued, presuppose a lawgiver or legislator. To give laws, a legislator must have legitimate authority over the moral agent. According to voluntarism, it is the will of a legislator with legitimate authority that imbues moral concepts with their normativity. Furthermore, the authority of the legislator stems from their ability to impose sanctions on those who break laws (Korsgaard, 1996, pp. 25-26).

There are many problems with voluntarism. The most obvious resembles the famous Euthyphro dilemma: What determines which moral concepts the legislator wills as a law?[15] If the legislator wills what is good because it is already good, then they are not the source of moral concepts, whereas if the good is only good because the legislator wills it, then the good is arbitrary. In response, Hobbes and Pufendorf argue that the legislature's will is not actually the source of moral content, only the normativity behind that content. They argue that morality concerns performing reasonable actions that allow humans to live together socially and the legislative authority merely establishes the normativity of these moral claims on us by giving us commands. The most fatal criticism of voluntarism involves questioning whether the authority of the legislator is in fact legitimate. If their source of authority is deemed arbitrary or only partial, then their ability to establish normativity is undermined.

The second historical solution presented by Korsgaard is the form of "moral realism" initially proposed by Clarke and Price (Korsgaard, 1996, p. 28). The authority of the legislative figure central to voluntarism can be easily called into question, and to avoid a similar problem, the realists aimed to establish an irreducible source of authority. They did this by arguing for the existence of intrinsically normative facts about morality, comparable to the other sorts of facts by which claims about the world are tested and verified. Thus, a moral concept could be considered normative if it were

[15] See footnote 13.

true, and it was true if it corresponded to moral facts about the universe. On this view, if, for example, “murder is wrong” is an intrinsic moral fact in the world, this would establish its normativity, and we should never murder. A serious problem arises for realism when we attempt to describe what these purported moral facts are. It seems their existence is difficult if not impossible to prove. The philosopher J. L. Mackie’s famously argued that if such properties existed, they would be queer, meaning they would be unlike any other property in the universe (Mackie, 1990). From the standpoint of the modern scientific worldview, this makes their existence highly dubious.

The third historical solution to the normative question was reflective endorsement. This is a descriptive term developed by Korsgaard to explain the source of normativity in the moral theories of Hume, Mill and Williams (Korsgaard, 1996, pp. 50-51). In each of these theories, normativity is established by appealing to an aspect of human nature. Because of our human nature, developed through evolution and enculturation, we find ourselves highly susceptible to moral claims on how we should act. For example, we have ethical intuitions and we can be moved by the suffering of others. This is often as far as the neuroscience goes, theorists such as Greene and Moll et al. stop at descriptions of these intuitions and their evolutionary origins. Once we have identified why we are moral, by describing our human nature, we can reflect on this nature and ask practical questions about whether it gives us a real or compelling reason to act. For example, we can ask, does this aspect of human nature, altruism say, generally encourage human flourishing, i.e. is it good for us? If altruism sufficiently answers our practical questions then we can endorse it, establishing its normativity. According to reflective endorsement the answer to “how should we act” arises from questioning our moral nature in practical ways and endorsing the aspects of our nature that we approve of.

A key problem for reflective endorsement theories concerns which practical questions should be asked and by which criteria should our moral nature be judged. Korsgaard’s solution, as we will soon see, similarly involves a form of reflective endorsement, however, its methods of generating reasons to act are different. In reflective endorsement theories, this criterion arises from preapproved goals or sentiments, i.e. the reason is preconceived and applied externally. Whereas in Korsgaard’s theory, the act of reflective endorsement itself generates reasons (Korsgaard, 1996, p. 89).

Korsgaard's own solution to the normative question is grounded in an account of individual autonomy, and synthesises the theories just summarised. The term "autonomy" originates from the Greek words "autos", meaning "self" and "nomos", meaning "law", literally meaning to "give laws unto oneself", i.e. to be "self-governing". It is also central to Kant's moral philosophy, as Kant equates autonomous action with moral action (Kant, 2002). Korsgaard (1996) argues that normativity arises from the process of reflecting on one's thoughts, identifying reasons to act, and by acting on the reasons that fit with what she calls a "practical identity". To choose in this way is to be a law unto oneself.

According to Korsgaard, voluntarism is at least partially true because normativity is established by the will of a legislator with legitimate authority. However, this legitimate legislator is oneself, not God or a despot. Realism is at least partially true because there are moral facts about the world, however they are not metaphysically queer facts that exist somewhere 'out there' in the universe, but rather they arise from facts about autonomous agency. Finally, reflective endorsement is at least partially true, because normativity is established by reflecting upon and endorsing one's thoughts as reasons, however there is no need to appeal to particular reasons to achieve this. As Korsgaard (1996, p. 89) puts it "…the reflective endorsement test is not merely a way of justifying morality. It is morality itself."

There are two major psychological assumptions required for Korsgaard's theory to get off the ground. The first is that humans have the ability to self-consciously reflect on the contents of their mind, e.g. their thoughts such as reasons and feelings. Secondly, the fact that humans are self-consciously reflective forces one to form a conception of oneself. Korsgaard calls this a "practical identity", e.g. being a student, or a New Zealander, as a way of self-understanding. "Practical identity" can be understood as a "description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking." (Korsgaard, 1996, p. 101).

Unlike most animals, which are forced to attend only to the world and have no ability for introspection, humans can turn their attention inwards and examine (at least in part) the contents of their mind. We can reflect on our desires, passions, reasons and intuitions. On the basis of this reflective process of the human mind, Korsgaard (1996)

formulates a distinction between the acting self and the thinking self. The capability of reflecting on one's mental activities means that one can "stand apart", as it were, from one's own thoughts. In this way, the thinking self can exercise power over the acting self by understanding the reasons why one acts, actively scrutinising these reasons, and by choosing which reasons to endorse. This is the process of acting in a deliberative way when making a decision or judgement. Humans appear to be the most developed of the animal world in this ability, however there is some evidence to suggest that our close ancestors, and certain animals, may also possess a rudimentary form of decision-making and reflection.[16]

The existence of this human capacity for self-conscious reflection presents the agent with a problem. If one can reflect on the contents of their mind and choose to act in a certain way, they are now faced with the dilemma of "how should I act?" (Korsgaard, 1996). One cannot simply ignore this question, as to ignore the question would require a conscious decision. One might become distracted and forget the question, but this would mean failing to act. So, when faced with a moral dilemma one is confronted by thoughts i.e. impulses to act, desires, passions, cognitive operations, all of which he or she must assess and ask themselves, are these really reasons to act? Therefore, the reflective nature of human consciousness is the origin of the normative question. Without it, one would have no ability to understand the reasons why one acts and exercise control over their actions, and the normative question would become nonsensical.

Upon reflecting on one's thoughts and desires, the outcome can be a rejection of these as potential reasons to act. For example, if I ask myself "should I give to charity?", I may notice that my desire to be altruistic is in fact based in a form of self-gratification. I can then ask myself, "is this a good reason to give to charity?" and perhaps conclude that it is not. Perhaps there may be something else I would rather do with my money. The reflective nature of human consciousness then disrupts the automaticity of action; it holds our acting selves back by rejecting our thoughts and desires as potential reasons.

[16] Hampton, R. R. (2001) 'Rhesus monkeys know when they remember', *Proceedings of the National Academy of Sciences*, 98(9), pp. 5359-5362. In this study, it was demonstrated that Rhesus monkeys had some insight into their memories. In other words, they were self-aware to an extent, about what information they could recall and which they could not.

For Korsgaard (1996) the solution to the normative question is implicit in the stating of the problem. As we have seen, reflecting on the contents of one's mind before acting can lead to the rejection of one's thoughts as potential reasons, and this is a problem. However, if some thoughts can be rejected then some may be endorsed. For Korsgaard "'reason' means reflective success" i.e. a desire, passion, feeling or cognitive operation that has survived one's introspective scrutiny and justifies a way of acting (Korsgaard, 1996, p. 97). To have a reason is to self-legislate, i.e. it is to answer the question of how we should act, and ascribe normativity to these claims on us. Such reasons are authoritative because they are our own and have emerged through successful reflection. Compare this to reasons that are hoisted onto us and that we do not endorse, which are only authoritative if they can be enforced. Even then, we can rebel. Finally, such reasons are a source of obligation, but they do not compel us to act. In deciding to do something one leaves open the possibility of doing otherwise. For example, if I need to get out of bed because I have work, I have a good reason to get up, but I do not have to. This is very different from being physically and forcefully dragged out of bed. In the moral sphere, if I have a good reason to donate to charity, I have an obligation as well, however I am not compelled to do so.

The reflective nature of human consciousness forces one to form a conception of oneself. And by reflecting on one's thoughts, i.e. deliberating, one endorses reasons by standing back from one's inclinations and choosing which to act on. This choice arises from oneself, and thus is contingent on one's self-understanding. As Korsgaard puts it:

> When you deliberate, it is if there were something over and above all of your desires, something which is you, and which chooses which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of yourself.
>
> (Korsgaard, 1996, p. 100)

In other words, the assessment of reasons arises from Korsgaard's second psychological assumption: that humans conceive themselves in terms of a "practical identity". Importantly, this "practical identity" is defined as a valuation of oneself and actions worth undertaking (Korsgaard, 1996, p. 101). Bundled into practical identities are reasons and obligations. For instance, part of my practical identity is that I am a

medical student, and there are certain obligations that arise from this. For example, I need to be careful not to discuss confidential health information in public. I have an obligation to learn, to become a competent doctor, so I must study hard. Many other obligations arise from other facets of my practical identity. Being a friend gives me reason to "check-in" with certain people. Being a New Zealander gives me reason to vote every three years. And so on and so forth. Hence, one's "practical identity" is the criterion by which one judges a thought as a "reason" to act.

An important aside is that practical identities are not necessarily separate, well-circumscribed or even well-conceived. Most individuals have a conception of themselves that more closely resembles a heterogenous mass, with facets of varying importance, which they use to anchor themselves to others and the world. Korsgaard (1996, p. 101) says: "Practical identity is a complex matter and for the average person there will be a jumble of such conceptions".

To illustrate Korsgaard's theory more fully, consider the following scenario: A medical student and her classmates are tasked with completing an assignment worth a substantial portion of their grade. The night before the due date, she observes two of her classmates copying each other's work. Plagiarising this assignment is a serious offence as it was designed to assess the competence of medical students in an essential skill for future clinical work. The medical student is now faced with a decision: she could either "turn a blind eye", confront her classmates about the nature of their actions, or report their plagiarism to the Dean of the medical school. On Korsgaard's analysis, the medical student acts by standing back and examining her thoughts, including her desires, intuitions, passions and any internal arguments. Through this process she can think through the various options and the implications of each, and then decide which of the potential reasons for acting to endorse. In the act of endorsement, she determines herself according to that reason. She may decide that a sense of loyalty or comradeship to her classmates is paramount, over and against her duty as a young professional. Alternatively, she may consider her responsibilities as a young professional and a trainee doctor, which includes upholding the integrity and standards of the profession. This will lead her to report the cheating. The option she takes then alters her "practical identity" according to the reasoning involved – it effects how she thinks through relevantly similarly situations in the future – and so defines who she is and how she values her life.

In the previous section, taking the example of Archibald Baxter, we examined Korsgaard's condition that any solution to the normative question must appeal deeply to one's identity (Korsgaard, 1996, pp. 17-18). If one does not carry out what they know they are morally obligated to do, it is corrosive to their sense of identity. For example, say the medical student, knowing that she has an obligation to report her classmate's plagiarism because she is a young professional, fails to do so. Or later in her career, suppose she spreads confidential information about a patient, and knows that this contradicts her moral obligations as a doctor. In so far as she actually cares and values herself as a good doctor, these failures will undermine her sense of self-worth. The self-contradiction will in turn destabilise any further reflections involving this aspect of her identity.

There is an obvious problem that arises by grounding moral obligation in individual practical identities. Moral obligations are meant to have a degree of universality, i.e. they cannot be dependent on what a person happens to be concerned about or find valuable. Murder, for example, is generally considered wrong for everyone, not simply wrong for those people who identify themselves as opposed to murder. This is a problem because some practical identities are completely arbitrary, inconsequential or even highly morally questionable. For example, a Viking 1200 years ago, reflecting on their identity, would have found reasons to loot, pillage and die in battle. These actions would have been quite consistent with their practical identity. Indeed, there is a popular image of an old distraught Viking who laments about not having died in their youth in a foreign land with sword in hand. Dying at an old age may be unstable to the Vikings identity. So, someone could act in ways consistent with their practical identity and yet be regarded as immoral or perhaps evil from the standpoint of others.

To overcome this challenge Korsgaard argues that rational reflection should lead us to a more fundamental identity, which she calls the identity of being human, or "the moral identity" (Korsgaard, 1996, p. 129). Though one can elect to change or reject aspects of one's identity, one cannot reject the need for identity altogether. Humans can self-consciously reflect on the contents of their minds, which forces them to have a conception of themselves in terms of a practical identity. This method of self-understanding is ubiquitous in humanity, so one can group all of humanity together by citing a need to form a practical identity. Korsgaard relates this to the universality

inherent in reasons. When something exists as a reason for one individual, it exists as a reason for another in a similar set of circumstances. For example, if it is wrong for one person to kill in particular circumstances, then the same applies to another person in the same circumstances. Therefore, when one reflects and endorses reasons in accordance with one's identity it is not a purely individualistic exercise. Humans obligate and make claims on each other based on shared reasons. Korsgaard argues that by appealing to this identity as a moral agent or an identification with humanity, one becomes conscious of certain obligations that are universal. This is similar to Kant's third conception of his categorical imperative, where he states that one should act as a member of the kingdom of ends (Kant, 2002). In identifying with humanity, one is conscious of an obligation to treat people as ends in themselves as opposed to means to ends (Korsgaard, 1996, p. 143).

Finally, the reliance of Korsgaard's theory on practical identity does not predispose it to critiques by "no-self" theories. Korsgaard does not propose a metaphysical model of the self, meaning that there is a thing (a self) that exists over and above one's mind and body. Instead she argues for the practical necessity of forming an identity (Korsgaard, 1989), and thus, avoids full-scale philosophical challenges of identity, like those made by the philosopher Derek Parfit. Parfit (1984) argues against the existence of a metaphysical self, stating that persons only consist of their:

> …brain and body, and the thinking of his thoughts, and the doing of his deeds, and the occurrence of many other physical and mental events.
>
> (Parfit, 1984, p. 275)

In Parfit's account, unity and integrity of the agent are still important, however, this can be achieved with a sense of psychological and physical connectedness with one's past selves (Parfit, 1984, pp. 301-302). By arguing for the practical necessity of an identity Korsgaard avoids this criticism. She argues that a unity of self is necessary for one to act in a coherent way. One may have many conflicting desires and thoughts, but one must decide upon a single way of acting and this process is unifying. She states: "your identity is in a quite literal way constituted by your choices and actions" (Korsgaard, 2009). Forming an identity is also necessary for an agent to act coherently across time. For example, long term goals only make sense from the point of an agent that identifies

with their future and past selves (Korsgaard, 1989). Therefore, Korsgaard's practical conception of identity avoids potential criticisms from "no-self" theories.

To summarise Korsgaard's moral philosophy so far: If one reflects on the contents of one's mind, one encounter thoughts (desires, passions and more cognitive operations) as possible reasons to act. One can either reject or endorse these mental contents as reasons and use them to self-legislate. By willing laws unto oneself, one creates reasons to act. The generation of reasons through reflective success is conducted by appealing to one's "practical identity". It is in this process of self-legislating, by reflecting and endorsing a reason in accordance with your identity which imbues certain ethical concepts with normativity (Korsgaard, 1996). In the following section I will discuss how these ideas relate to the neuroscience of moral cognition.

3.2 Potential Contributions of Korsgaard's Moral Philosophy to Moral Cognition

In this chapter so far, I have outlined Korsgaard's account of moral obligation. This account is her answer to the normative question: the question of where the sense of "ought", inherent in all moral claims, comes from. In this section I will further elaborate on how Korsgaard's theory might advance the neuroscientific study of moral cognition.

In a recent review of the literature of moral cognition (focusing on the use of EGG) from a bioethical perspective, Wagner et al. (2017) expressed the need for more theoretical work, specifically to examine the relationship between the field's philosophical and neuroscientific facets. Korsgaard's account of moral obligation, centred on the integrating role of reason in human agency, will contribute to this kind of theoretical work in multiple ways: Firstly, Korsgaard's theory of moral decision-making is broadly consistent with the nature of an integrated account of moral cognition. Her theory emphasises agential unity, which is the integration of multiple underlying processes to achieve a singular goal or carry out a single directed operation. Korsgaard claims that one reflects and endorses an inclination as a reason to act. This process is necessarily unifying, as any underlying processing must culminate in a single directed decision. The literature suggests that the brain regions and processes involved in moral cognition operate as a network and are thus in some sense unified. Secondly,

Korsgaard's theory reveals a critical link between reason and reflection, that has been largely ignored in the contemporary neuroscientific literature, and yet is highly suitable for neuroscientific study. Thirdly, by setting out three conditions that must be met to answer the normative question, Korsgaard's theory indicates inability of neuroscience to describe the normative aspects of morality. In the next section I will demonstrate why exactly neuroscience cannot answer the normative question, which demonstrates why moral philosophy is required alongside neuroscience to describe the relationship between morality and the brain.

As Korsgaard's theory primarily attempts to give a compelling account of the normative aspects of morality, grounded with psychology, and is broadly consistent with an integrated account of morality, it is an ideal candidate for an 'off-the-shelf' theory in moral philosophy, to utilise in the formulation of an integrated account of moral cognition.

3.2.1 The Normative Question and the Neuroscientific Study of Moral Cognition

In the light of Korsgaard's three conditions for answering the normative question,[17] it is worth considering how it might be answered by neuroscience, and how it compares to the questions neuroscientists typically try to answer. One difference appears to be that while philosophers like Korsgaard are looking to justify the normativity inherent in moral claims, neuroscientists are typically seeking a descriptive account of human moral cognition. They use the scientific language of psychology, cognitive neuroscience and evolutionary biology to present a "detached" view of the relevant behaviour. Moreover, this detached view usually takes the form of a causal story, in the same way that a physicist might explain why an apple fell from a tree. In this respect, they are not actually attempting to answer the normative question at all, because they are not meeting Korsgaard's first criteria. One might say that whereas the neuroscientist attempts to explain why we are moral, the philosopher is attempting to say why we should be moral.

[17] To summarise, a theory must address us from a first-person perspective, be transparent, and appeal deeply to our sense of identity. Korsgaard, C. (1996) *The sources of normativity.* Cambridge University Press. pp. 17-18

To further explore this comparison, let us consider how two "working definitions" of "Moral Cognition" currently being used in the neuroscientific study of moral judgement might apply to the normative question. As explained in chapter two, definitions such as these are methodologically necessary to define the borders of the empirical search for the diverse cognitive faculties that should be regarded as part of human moral cognition. An example is Moll et al.'s (2008), which is focussed on the motivational factors underpinning social behaviour. According to this definition, if the motivation behind an action or decision was that of "genuine altruism" then it was relevant to moral cognition, and this is what distinguished it from other kinds of social behaviour. Another example is Greene's definition, which focusses on the evolutionary function of specific kinds of behaviours. According to Greene (2014), moral cognition is concerned with studying psychological adaptations that functioned to enhance group co-operation in otherwise selfish individuals.

Moll et al. and Greene represent typical approaches in the modern neuroscientific study of moral cognition, as most existing attempts at defining moral cognition in neuroscience focus on explaining either the specific evolutionary and developmental origin of human moral thinking (adaptations to enhance co-operation according to Greene (2014)) or the specific psychological origin of human moral thinking (genuine altruism according to Moll et al. (2008)). In other words, the focus is on describing the instantiation and features of moral cognition. Such descriptions might enable us to identify patterns of thinking and behaviour, casual relationships behind these patterns, the brain processes underlying these behaviours and more fundamentally their evolutionary or developmental origin. While there is value in this work in explaining the processes behind human moral thinking, it is important to recognise how it is not addressing the normative question. This becomes clear when we apply Korsgaard's three conditions to Greene and Moll et al.'s definitions. To illustrate, suppose I were to ask myself "why should I give to charity?", as a personal question (the first criteria). If one looks to Moll et al. (2008) for an answer it might be this: "I give to charity because I am motivated by a deeply ingrained sense of genuine altruism". From a personal standpoint, this answer fails to adequately address the normative question. By reflecting on such a motivation and scrutinising it, one realises that it is not inherently motivating, and it can easily be overridden by a different motivation. Moll et al.'s definition does not address this potential moral conflict.

According to Korsgaard only through reflection can one tell whether their altruistic motivation is well placed.

If I look to Greene's definition for an answer, I find something like this: "because this behaviour arises from the psychological adaptation of altruism that enhanced group co-operation in my ancestors, which through the evolutionary process of natural selection became a deeply ingrained psychological factor" (Greene, 2014). It may be that I have a sense of altruism because it helped my ancestors survive and reproduce, but now that I can reflect, that does not mean that in this instance I ought to give to charity. While these definitions attempt to explain the causality behind morality, they do not provide an explanation of the authority of moral claims. Merely appealing to these answers alone would not justify acting on any moral duty from a personal standpoint. Many neuroscientists are aware of the difficultly applying neuroscientific theory to moral philosophy. For example, Greene suggests a further answer to the normative question. He does this by admitting that we may begin to question whether we are truly beholden to our evolutionarily derived moral instincts, and that our moral inclinations may not survive this knowledge (Greene, 2014, p. 25). He is particularly aware that having a bias and preference towards your in-group is a perfectly viable psychological adaptation that would enhance in-group co-operation, however it may also be a source of racism and xenophobia. In "Moral Tribes", Greene (2014) illustrates this dichotomy between an explanatory moral theory and a normative moral theory, by showing how human morality can extend beyond its evolutionary origins:

> …we can take morality in new directions that nature never 'intended'. We can, for example, donate money to faraway strangers without expecting anything in return. From a biological point of view, this is just a backfiring glitch, much like the invention of birth control. But from our point of view, as moral beings who can kick away the evolutionary ladder, it may be exactly what we want. Morality is more than what it evolved to be.
>
> (Greene, 2014, p. 25)

Evolution might explain why people are unreflectively moral, but it does not justify the claims morality makes on us, in a way that is likely to survive personal reflection. Interestingly, Korsgaard uses a similar example to show how a particular moral theory could have sufficient explanatory adequacy (illustrating where human morality comes

from) but lack normative adequacy (a solution to the normative question) (Korsgaard, 1996, pp. 14-15).

Greene is a utilitarian, and he arrives at this normative conclusion through a philosophical analysis alongside a review of the neuroscience. The ethical beliefs of Korsgaard, a neo-Kantian and Greene, are very different, yet Greene agrees with Korsgaard to the extent that he recognises that morality must extend beyond any neuroscientific or psychological description of its instantiation. Korsgaard has constructed her own modern solution to the normative question, and Greene has done something very similar by giving his answer to "how shall we act?" (Greene, 2014). Greene appeals to the standard utilitarian principles, which are vulnerable to the standard objections. To properly account for moral cognition, it is necessary to go beyond mere explanations of why we feel we have specific moral obligations and explain whether and why we are obligated at all, knowing the origins of those feelings.

The fundamental distinction between the explanatory and the normative is that the former, broadly using neuroscience, describes morality from a causal standpoint, whereas the latter, describes morality from a personal agential standpoint. Despite this distinction between scientific and philosophical accounts of morality, there are some important points of intersection. It is undeniable that humans can be deeply moved by a sense of moral "ought". So, at a basic level neuroscience-based models of moral cognition should be able to account for the fact that ethical concepts appear normative to us. Conversely, a philosophical theory answering the normative question that did not fit with current neuroscience would for that reason be suspect.

If one wants to understand the relationship between the brain and our moral lives, and neglects an account of normativity, one is missing something integral to human moral functioning. If one wants to explain moral cognition in its entirety, relying on only a normative or an explanatory model will render an incomplete picture. Thus, an interdisciplinary and integrative account of moral cognition requires both normative and explanatory aspects.

3.2.2 Importance of Reason and Reflection in Moral Cognition

In chapter two I described both the modern and historically predominant models of moral cognition. In the 1970's, rationalist models like those proposed by Kohlberg (1971) were popular. These drew on the Kantian philosophical tradition. The important

characteristic in Kohlberg's definition of morality is the formal character of moral judgements as opposed to their contents. On this view, the formation of a moral judgement is more important than the moral prescription. In this respect Kohlberg's model resembles Korsgaard's theory, which identifies the formal character of reasons as the basis of moral obligation. Kohlberg states:

> Impersonality, ideality, universalizability, preemptiveness, etc. are the formal characteristics of a moral judgment. These are best seen in the reasons given for a moral judgment, a moral reason being one which has these properties. But we claim that the formal definition of morality only works when we recognize that there are developmental levels of moral judgment which increasingly approximates the philosopher's moral form. This recognition shows us (a) that there are formal criteria which make judgments moral, (b) that these are only fully met by the most mature stage of moral judgment, so that (c) our mature stages of judgment are more moral (in the formalist sense, more morally adequate) than less mature stages.
>
> (Kohlberg, 1971, p. 215)

In Kohlberg's account of morality, in the fullest form of human moral development reasons have the above four features (Impersonality, ideality, universalizability, preemptiveness), in other words reasons must be what we could characterise as "objective".

In the late 1990's and early 2000's Kohlberg's rationalist theories were widely rejected in the mainstream in developmental and moral psychology, on account of their excessive focus on "objective reasoning", and corresponding neglect of emotions and desires in forming moral judgements. As an alternative, Haidt developed his "social intuitionist" model of moral judgement, where a moral judgement is formed by intuitive and emotional reactions, with reasons being sought to justify the judgement afterwards (Haidt, 2001). This model presents an important challenge for Korsgaard's moral philosophy, which I will discuss in detail later in the chapter.

Sentimentalist models of moral cognition like Haidt's became increasingly popular in the twenty-first century. Consequently, the role of reason and reflection has been largely side-lined in research into moral cognition, with much of the focus being

placed on emotion. To a certain extent this may be appropriate. However, the increased attention on automatic and intuitive judgements has left the reasoning and reflecting aspects of moral judgement underdeveloped. Moreover, those models do involve reasoning, like the "dual-process" model proposed by Greene (2016), still feature fundamentally sentimentalist assumptions. This is because the debate about the relationship between reason and emotion, ongoing in Haidt and Greene's work, maintains what seems is a false dichotomy between the two. As I will discuss in the next chapter, and what was briefly mentioned concerning an integrative neuroscience in chapter two, there are compelling reasons to think that reason and emotion are not unrelated and opposing forces, as the various proponents in this debate generally assume. Haidt and Greene both hold that generally moral judgement are either made through reasoning, or through emotive intuitions, but not both. The primary difference between these two theorists is that Greene thinks sometimes, given certain contexts and conditions reason can be employed in moral judgements, whereas Haidt argues for the primacy of emotion in all contexts (Greene, 2016; Haidt, 2001).

Korsgaard's theory centres on an account of reason and reflection (Korsgaard, 1996). This focus is contrary to the current trends in the literature, and thus her theory can contribute to the development of a neglected areas in the study of moral cognition. Furthermore, Korsgaard's view of practical reasoning further highlights the dubious dichotomy between reason and emotion, which emerges from the view of moral cognition as a network. According to her theory emotions, feelings, passions, empathy, sympathy can all be potential "reasons" to act, if they are endorsed through reflection. Through an analysis of the literature we saw in chapter one, that moral cognition is multi-faceted, involving emotion, reason and social-cognition, which is reflected by Korsgaard's moral philosophy.

3.2.3 Response to Challenges Facing Korsgaard's Psychological Assumption of Self-Conscious Reflection

In chapter four I will compare Korsgaard's theory with findings from neuroscience. Beforehand it is necessary to critically examine the key psychological propositions involved in her theory, from the standpoint of moral psychology. Moral psychology is the study of human behaviour and thought in moral contexts. It is necessarily interdisciplinary, integrating psychology, neuroscience and moral philosophy.

As I discussed earlier in this chapter, there are two major psychological assumptions that are an essential grounding for Korsgaard's theory (Korsgaard, 1996). The first is that humans have the capability of self-conscious reflection. This is the idea that humans can reflect on (at least part of) the contents of their minds. The second assumption is that this self-conscious reflection forces one to form a conception of oneself, and this process of self-understanding involves notions of value and purpose (i.e. a self-description that Korsgaard refers to as a "practical identity"). Korsgaard's reliance on psychological assumptions illustrates the inherent interdisciplinarity of moral cognition. While she gives a normative account of morality, psychology and neuroscience are still essential. This is because Korsgaard's account of reason and reflection emerge from two psychological facts and are thus beholden to neuroscience. If for example, reason and reflection are shown to be subservient to emotion during moral judgements, this would undermine her theory. Thus, a philosophically informed neuroscience will also set reasonable limits to any normative account of morality.

Both of Korsgaard's assumptions are uncontroversial on the surface. It is obvious that people identify with certain groups, which provides one with a sense of value and purpose. Ask anyone you meet to describe who they are, and they will describe themselves in various ways that constitute a practical identity. Someone may define themselves in terms of their nationality, religion or lack thereof, their gender and their profession. As discussed earlier, Korsgaard (1996, p. 101) describes one's personal identity as a "jumble", a complex group of conceptions, which range in importance to the individual. One may not reveal themselves so directly, but practical identities are an essential way of understanding oneself as an agent. Secondly, it is obvious that humans can reflect on the contents of their minds, allowing one to understand one's reasons for acting. Humans are not completely blind to their own inner workings, and more often than not can give their reasons for acting in various ways.

As a reader, you can do this now by simply reflecting on your feelings or thoughts about my arguments so far. Introspection is uncontroversial; however, many of one's thoughts and reasons can be obscured. Individuals can be oblivious of their desires and biases that lead them to act, and upon reflection one may be uncertain why one acted a given way at all.

To test the role of reason in establishing moral judgements the social psychologist Jonathan Haidt (2001) formulated an experiment called the "dumfounding scenario". Based on this experiment Haidt developed the "social intuitionist model" of moral judgement in his famous article "the emotional dog and its rational tail" (Haidt, 2001). The test involves presenting a scenario that includes various cultural taboos to a participant and asks them to make a judgement about the rightness or wrongness of the actions depicted. Haidt's article begins by giving an example of such a scenario:

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love?
>
> (Haidt, 2001, p. 814)

Haidt (2001) discusses how most participants say what was depicted was wrong, and then they are asked to give reasons. However, as you can see in the example, cunningly built into the scenario are adequate responses to the common reasons given by most participants. For example, they point out the potential problems with inbreeding, and the harm that may be caused. However, in the scenario the siblings used contraception and it explicitly states that no one was harmed. Finally, Haidt (2001, p. 814) says, participants say something like '"I don't know, I can't explain it, I just know it's wrong."'

If it were shown that reflection had little to no effect on one's reasoning or moral judgements, then this would be a serious problem for Korsgaard's theory. The dumfounding scenario however was employed by Haidt to explicitly critique rationalist models of moral judgement, by claiming that reasoning is post-hoc and has little effect on the outcomes of moral judgements. In the excerpt above, participants seemingly searched for reasons when the interviewer inquired after them. So, the initial judgment appears pre-reflective. The interviewer then proceeded to take the participant through a guided reflection, and the reasons employed by participants were always met with

counter-reasons. Despite all the arguments against any "reason" the participant could wield to claim the actions presented were wrong, their initial judgement was conserved. Haidt concludes by presenting his "social intuitionist model of moral judgment" where primarily only initial intuitions influence judgments. Importantly Haidt (2001) acknowledges that occasionally private reflection and observing a peers reasoning may alter or inform one's judgements, but he argues this is much rarer than rationalist theories like Kohlberg claim.

Haidt's conclusion is not simply problematic for Korsgaard's theory, but for any theory of autonomy, as it entails that one has little control over one's judgements. However, it is not clear that Haidt's observations justify his conclusions. Common sense seems to show that they do not. Consider your own experience and intuitions concerning reflection. Can you recall a moment recently where facing a difficult decision you had an urge to act one way, but then made yourself reflect which changed your mind? Just over a year ago I became vegetarian. Beforehand, I remember operating off an intuitive judgement that consuming meat was permissible, and despite regularly encountering good reasons to be vegetarian, they barely affected me. In the months before making the change I became internally hostile to the idea of vegetarianism, and occasionally I would find myself silently condemning vegans and vegetarians who I saw as "preachy" and militant. Finally, over the course of a week or so, upon reflecting again on the arguments for vegetarianism I made the lifestyle change. So, through the process of rational reflection, I was forced to acknowledge an imperative not to eat meat. Having made this decision, it is now the case that continuing to eat meat would be at odds with this identity. In retrospect I understand that the couple months of increased hostility towards vegetarianism was a manifestation of the internal struggle between my identity and my failing actions.

There are reasons to doubt Haidt's methodology. The kinds of dilemmas he employs are highly specific, as they concern moral judgements about "purity", such as an intrinsic condemnation of incest. Purity judgements are absolute judgements against certain acts deemed profane, disgusting and abhorrent. Perhaps human reflection is weakest against intuitive judgements of this kind, because they are more deeply ingrained into the human psyche (and perhaps this is an evolutionary reason). However, this does not mean that reflection cannot be a powerful tool to change many other kinds of judgements, such as harm judgements. Take arguments for vegetarianism as an

example. Even if a person is unconvinced by the arguments for vegetarianism, one would nevertheless expect them to give some reason for their continued consumption of meat (this could even be that "it tastes good and I value my happiness over animals"). This raises questions about the generalisability of Haidt's findings (Haidt, 2001), which may only apply to certain kinds of purity judgements, which are often steeped in culture.

Finally, there is experimental counter-evidence which shows that reflection does influence one's reasoning and moral judgements. A study by Paxton et al. (2012) showed just this, using Haidt's own "dumfounding scenario" to show that reflection does alter how one responds to reasons and makes judgements. Their experiment had two independent (manipulated) variables, the first was argument strength and the second was temporal duration before a judgement. Participants were given Haidt's own "Julie and Mark incestuous relationship" vignette that I introduced earlier in this section. Following this, half of the participants were presented with a "Strong argument" for why Julie and Mark's actions were permissible, designed to be as persuasive as possible, and the other half were given a "Weak argument" for why their actions were permissible, which was designed to be poorly persuasive. Following this, half the participants from each group were asked to judge the permissibility of Julie and Mark's actions immediately, while the other half were given an additional two minutes to think while the argument remained on the screen. Permissibility scores, the "moral acceptability rating" were rated on a scale of 1-7, with 7 being the most morally acceptable and 1 being the least. When reflection was not encouraged (i.e. the judgement was made immediately after being exposed to the arguments) argument strength had no effect on the permissibility rating given by participants, and in both cases (weak and strong argument groups) the mean moral acceptability rating was rated at around 3. However, when reflection was encouraged, participants exposed to a strong argument rated the moral acceptability as much higher than the group exposed to a weak argument (with a mean rating of around 4 vs. 2 respectively). This shows that reflection not only altered the judgements of the participants, but it was necessary for argument strength i.e. reasons to influence the judgement (Paxton et al., 2012).

It seems then that Korsgaard's psychological assumptions concerning the human ability to self-consciously reflect to endorse reasons and her appeal to practical identity, can withstand deeper psychological and philosophical scrutiny. I have not

addressed all potential criticisms of Korsgaard's moral philosophy, but I have shown that it can survive these common attempts at refutation.

3.3 Conclusion

In this Chapter I presented Korsgaard's neo-Kantian theory of moral obligation, which attempts to answer the "normative question" (Korsgaard, 1996). I showed that explanatory accounts of morality, typical in neuroscience and psychology, cannot answer the normative question. Hence, an integrative account of moral cognition, with the goal of adequately explaining the relationship between the brain and morality, requires both neuroscience and moral philosophy to explain both explanatory and normative accounts. I also discussed how Korsgaard's moral philosophy might benefit the field of moral cognition. In accordance with a need to carry out an interdisciplinary analysis within moral cognition and to further develop theoretical work in the field, aspects of Korsgaard's theory will be used. In particular, the emphasis on reason and reflection which have recently seen limited focus in moral cognition. Finally, from a neuroscientific standpoint, Korgaard's moral philosophy resonates with much of the evidence in the literature and can withstand critiques of identity from philosophers like Parfit and critiques of reflection from psychologists like Haidt.

So far in this thesis I have discussed the need for an interdisciplinary and integrative account of moral cognition and have done so in unidirectional way. I have used moral philosophy to critique neuroscience, to show the need for the three features of an "integrative account". In the next chapter I will change directions and attempt to integrate the aspects of Korsgaard's moral theory focused on reason and reflection with other perspectives in the neuroscience. I will introduce neuroscientific perspectives from the literature reviewed in chapter one and the work of the neurologist Antonio Damasio to critique her theory. This process is an exercise in interdisciplinarity, the goal of which is to modify Korsgaard's theory so that it better reflects the neuroscience. This process of modification and refinement will act as the interdisciplinary work in laying out the framework for an integrative model of moral cognition.

4 The Framework of an Integrative Account of Moral Cognition

The overarching objective of this thesis is to discuss the need for an account of moral cognition that is integrative in its character and interdisciplinary in its approach. In chapter one, through a select review of the literature, I laid out the background of the neuroscience of moral cognition. The empirical evidence suggests that moral cognition involves diverse brain regions, disparate in space and function. These regions operate as a network by building off their respective domain general functions. In chapter two I discussed the challenges facing the neuroscience of moral cognition. These challenges indicate the inadequacy of neuroscience alone in describing the relationship between the brain and morality, and the need for an interdisciplinary analysis, that includes social psychology and empirical neuroscience, alongside theoretical neuroscience and moral philosophy. A key part of the interdisciplinary work should be developing an account of how brain processes are integrated in moral cognition. Such a focus requires the use of theories of higher-level cognition.

In chapter three I began this interdisciplinary analysis by introducing the work of Christine Korsgaard (1996) as a philosophical framework through which to consider moral cognition. The chapter began by introducing what Korsgaard describes as "the normative question", i.e. the question of why we "ought" to follow the requirements of morality. An answer to this question is necessary to properly explain morality. The other component, an explanatory account of morality, focuses on a casual description of morality, typical to neuroscience and psychology. In chapter three, I showed that explanatory accounts of morality could not answer the normative question. Hence, an integrated account of moral cognition addresses both explanatory and normative questions.

In chapter three I also responded to a potential criticism of Korsgaard's theory. This criticism targeted her psychological assumption that humans are self-consciously and rationally reflective, and that it is this reflection that determines our moral judgements. I showed that rational reflection likely does play an important role in our moral lives, and that these faculties are often deemphasised in the current neuroscientific literature of moral judgement.

The framework described in this thesis has been called an integrative account of moral cognition. In the introduction I laid out its features, which are all "integrative" in some sense:

1. Neuroscientifically integrated: Moral Cognition involves integrative brain processes and as such should be described using theories of higher-level cognition.
2. Integrated across disciplines (as in "inter-disciplinary"): From a methodological standpoint any comprehensive description of moral cognition ought to refer to philosophy, neuroscience and psychology and (at least) integrate their respective methods, generating a synthesis of perspectives.
3. Integrated explanatory and normative accounts of morality: To form a complete account of moral cognition, both explanatory and normative descriptions are required.

So far, in this thesis I have travelled down a single road, in discussing the need for an integrative and interdisciplinary account of moral cognition. This discussion has centred around demonstrating the need for the three features above, so that one can describe the relationship between the brain and morality. Now, I will change directions, and carry out the interdisciplinary analysis. So far, I have largely used moral philosophy to critically analyse neuroscience, now is the time for some reciprocation.

In this chapter I will formulate a model of moral cognition based on this "integrative account". This framework will be developed by critically analysing Korsgaard's theory from a neuroscientific perspective, primarily using the work of Antonio Damasio and key cognitive neuroscience studies that featured in chapter one.[18] There are two important issues that this framework attempts to resolve. The first issue is the apparent distinction between automatic and animalistic (i.e. non-autonomous) ways of "acting", and highly deliberative and reflective acting. In Korsgaard's theory, this appears to be a hard distinction, with no room for gradations. Using Damasio and the neuroscientific literature I will argue that this dichotomy is false, and that there is a spectrum of automaticity and deliberation in moral judgements. The second issue is the

[18] Damasio broadly lays out his theory of decision-making in: Damasio, A. (1994) *Descartes' error: emotion, reason, and the human brain.* G. P. Putnam's Sons.

dichotomy between emotion and reason. I will also discuss why this is problematic and show that both faculties are intertwined and underlie practical reasoning.

4.1 Damasio's Somatic Marker Hypothesis

A major hurdle for models of moral cognition in which reflection and reasoning have a primary role is in the explaining how reasons can motivation. As mentioned in chapter two, this challenge was crystallised by Hume:

> We speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them.
>
> (Hume, 2007, p. 266)

For Hume, and his modern exponents such as Prinz, reasons do not motivate; only passions or sentiments do. In their view, when someone is faced with a moral dilemma, there are not competing reasons for different ways to act (e.g. the rationalists), nor is reason competing against emotion (e.g. Greene's dual process theory), but rather two opposing sentiments compete to motivate action. The role of reason, on these theories, is to enable us to think through the conflicting sentiments, and to compare their implications. Prinz is a sentimentalist and a motivational internalist (Prinz, 2007). The former means he thinks that sentiments constitute moral judgements, and the latter means he thinks that moral judgements necessarily motivate. These two positions go hand in hand, because if sentiments (a feeling or emotion) intrinsically motivate actions, and they are a judgement of right or wrong, then moral concepts are always linked to a motivation of some kind (Prinz, 2007, pp. 18-19). Prinz gives some compelling evidence for the necessity of emotions in motivation. Prinz (2007) argues that all moral judgements and actions (unless you are a psychopath) are accompanied by emotions. It is impossible to act in accordance with a moral judgement without feeling, and it is impossible to do something one knows is morally wrong (like stealing even if it is advantageous to oneself) without feeling a "bad" feeling, such as shame, guilt, sadness or remorse. This claim, that an emotional response of some kind is necessary to motivate moral judgements, seems in direct opposition to an account of

morality built on reason and reflection, such as Korsgaard's. However, the relationship between reason and emotion is far more complex than what is generally assumed.

Evidence on the necessity of feelings to guide and motivate action was gathered in an extensive study led by the neurologist Antonio Damasio in the late 1980's and early 1990's (Damasio et al., 1990). Damasio studied patients with a specific grouping of brain lesions, involving the ventromedial prefrontal cortex (vmPFC).[19] Damasio made several key observations of these patients. While they maintained their intellect, their decision making was greatly disrupted, largely concerning the personal and social realms (Damasio, 1996). Before the onset of the brain lesion, Damasio described these patients as "intelligent, creative and successful", however now their ability to plan and organise their life was greatly disturbed. They would make choices that would lead to financial ruin, the disintegration of relationships, and which were generally not "personally advantageous". Furthermore, their decisions were substantially different from what they would have been before developing the brain lesion.

In the light of these observations, Damasio (1996) formulated the "somatic marker hypothesis". The hypothesis states that emotions or feelings are a necessary "tether" for someone in making decisions, guiding and constraining reasoning, and motivating them to act appropriately, morally, socially and pragmatically (Damasio, 1996). Throughout life, one encounters and learns from various practical and social situations and problems. In these contexts, one begins to associate particular feelings with actions and outcomes. Damasio refers to these feelings as "somatic markers", meaning that they are embodied feelings ("gut feelings") that include visceral and nonvisceral sensation (Damasio, 1994, p. 173). These somatic markers are subsequently re-experienced when encountering similar situations or contexts. Certain somatic markers would become associated with specific facts, experiences and actions, which are valued to varying degrees as either positive or negative. This allows for the reactivation of these feelings to help constrain potential reasoning, by alerting people to the "goodness" or "badness" of certain actions and their corresponding outcomes in a shared context of action. Now, instead of needing to reason through an extensive cost-benefit analysis for every decision, these somatic markers can deter one from even

[19] As outlined in Chapter One, the vmPFC is also the brain region damaged in the famous Phineas Gage, who according to Henry Harlowe, presented with similar social deficits to Damasio's patients. Damasio uses Gage as his first example in his book ibid.

considering potentially disastrous actions by associating them with extremely negative feelings. This process constrains the range of options one considers, allowing one to make decisions quickly, efficiently and more effectively. Relying on only reasoning, without any somatic states to guide one's thoughts can be overwhelming.

To illustrate his theory, Damasio (1994) tells the story of one of his brain lesion patients who recently recounted a traffic accident involving road ice that he had witnessed, and now struggled to decide between two dates for his next laboratory visit:

> For the better part of a half-hour, the patient enumerated reasons for and against each of the two dates: previous engagements, proximity to other engagements, possible meteorological conditions, virtually anything one could reasonably think about concerning a simple date. Just as calmly as he had driven over the ice, and recounted that episode, he was now walking us through a tiresome cost-benefit analysis, an endless outlining and fruitless comparison of options and possible consequences. It took enormous discipline to listen to all of this without pounding on the table and telling him to stop, but we finally did tell him, quietly, that he should come on the second of the alternative dates. His response was equally calm and prompt. He simply said: "That's fine." Back the appointment book went into his pocket, and then he was off.
>
> (Damasio, 1994, pp. 192-194)

Without any guiding or motivating feeling the patient was overwhelmed with the potential avenues of actions and reasons ad infinitum to process. This illustrates the necessary merger of reasoning and feelings in order to function practically at all, reflecting Damasio's claim that a "[r]eduction in emotion may constitute an equally important source of irrational behaviour" (Damasio, 1994). Feelings appear to be essential parts of practical reasoning.

To explore the evidence supporting "somatic marker hypothesis" let us take two examples that demonstrate the deficits common to vmPFC patients: In his clinical work as a neurologist, Damasio encountered a patient he identified as sharing many of the same symptoms as Phineas Gage – the 19th century railroad worker turned famous neuropsychology patient.[20] In Damasio's book, "Descartes' Error" he refers to this

[20] Refer chapter one for a summary of his story and symptoms.

patient as Elliot, who he first met as a man diagnosed with a meningioma (a benign tumour that can impinge upon underlying brain tissue) (Damasio, 1994, p. 35). After first impressions, he noted Elliot was "pleasant and intriguing, thoroughly charming but emotionally contained", "coherent and smart", a "good husband and father" and a successful businessman (Damasio, 1994, pp. 34-35). Elliot underwent surgery to remove the tumour, however, soon afterwards Damasio noticed some lasting effect that the surgery had on his personality. He began to act in ways that were eerily similar to Gage.[21] At work Elliot managed his time poorly and was unable to keep track of the "overall purpose" of his work, becoming consumed by peripheral and lengthy tasks. On an intellectual level he knew what work needed to be done and what procedures to use, but his decision-making was disturbed, and he would spend whole afternoons deliberating about minor tasks (Damasio, 1994, p. 36). Elliot lost the trust of his co-workers and his job was terminated. Soon afterwards he started up new businesses, associated with a "disreputable character", made foolish financial decisions and went bankrupt. He was divorced from his wife, married a second time, and was divorced again soon after. Elliot had intact IQ, performed normally on memory tests, and even performed well in the laboratory responding to tasks concerning social convention and moral value (Damasio, 1994, pp. 40-46). But in practice, when making important decisions out in the world of shared human activity, he failed again and again.

Damasio included Elliot in an experiment where participants were presented with emotionally charged stimuli (e.g. burning houses), and their responses recorded. Elliot informed him afterwards that before the accident he would have reacted greatly to these kinds of pictures, but they no longer caused him to feel anything. This lead Damasio (1994, p. 45) to surmise his situation as "[t]o know but not to feel". Similar to Phineas Gage, it was the vmPFC which was lesioned in Elliot, hence their striking similarities (Damasio et al., 1994; Damasio, 1994).

The second example is experimental and involved tracking the behaviour and emotional responses of patients with vmPFC lesions. Bechara et al. (1996) studied the skin conductance response (SCR), a measurable physical response that indicates anxiety or arousal, of controls and vmPFC lesion patients when they underwent a risky

[21] Elliot, being a contemporary case means that any clinical observations made about him compared to Gage are more empirically useful. Although Gage's story is captivating, an overreliance on his case can be dubious, as his presentation can be susceptible to retroactive embellishment.

decision-making task. Subsequently named the Iowa Gambling task, it involved placing four decks of cards in front of the participant and giving them a fictitious loan of $2000 to gamble with. Participants were then asked to repeatedly draw cards from any of the four decks, A, B, C and D with the goal being to make as much money as possible. Some cards in each deck caused the participants to win money, while others caused them to lose money, with the compositions of the decks being different. Decks A and B were high risk decks, the participants could potentially draw cards worth $100, but in the long term, because of higher values of penalty cards they would net lose money. Decks C and D were safe decks, the participants could potentially win only $50 with each draw, but the penalty cards were less and over the long term the participant would net win money. The compositions of the decks were initially unknown to the participants. Both the control and lesion groups exhibited SCRs after choosing a card, indicating an emotional response while experiencing either an award or punishment. However, over the course of the experiment, only control participants exhibited an anticipatory SCR, that occurred before drawing a card, and this response was greater before drawing cards from the risky decks. Lesion patients exhibited no such anticipatory SCR. In a similar experiment using the Iowa Gambling task it was shown that control groups shift towards picking from the advantageous decks over the course of the experiment, while the lesion group continued to pick from the disadvantageous deck (Bechara et al., 1994). Analysing these two experiments together, Bechara et al. (1996) hypothesised that the anticipatory SCR, indicating the presence of somatic markers, guide the participant's decision-making by attributing positive or negative values to each deck. Patients with vmPFC lesions continually make disadvantageous decisions because of this absent biasing in potential choices. There is no negative somatic marker stopping the lesion patients for making dangerous or risky decisions. This might explain why Elliot made such poor decisions, and perhaps why Gage's personality changed so drastically.

One might take Damasio's work as reinforcing the Humean thesis maintained by Prinz and others, in that his conclusion shows that feelings are necessary to act coherently in social and moral contexts. However, it actually shows that reasons and feelings are intertwined, and that both are necessary for one to navigate through the world practically and socially. Reasoning is still employed to categorise and direct actions, however it is guided and constrained by feelings – together they constitute a

practical form of reasoning. This dispels the arguments for the primacy of either emotion or reason, and instead presents a more complex model of decision making. In this way Damasio's work is congruent with the neuroscientifically integrative feature of an integrative account of moral cognition that was discussed in chapter two.

Damasio's hypothesis also raises many serious questions for Korsgaard's moral theory. Her account of moral motivation, relying on endorsing reasons by appealing to one's identity, is problematic given Damasio's hypothesis. According to the "somatic marker hypothesis" feelings and emotions are essential motivating and guiding factors in reasoning (Damasio, 1994). If one takes a traditional Kantian view of reason, that it is an objective, fact-based justification for acting, and that reason alone motivates, this is incompatible with Damasio's account of morality. If one must endorse an inclination as a reason to act, and this reason alone is sufficient to motivate and guide actions, then why do we see so many problems with decision making in Damasio's brain lesion patients? In these patients the link between emotions and decisions is severed. However, Korsgaard defines reason merely as "reflective success" (1996, p. 97). Reasoning, emotions, feelings, empathy and inclinations can be constituents of a moral judgement and be endorsed as reasons. With this in mind, one can see Korsgaard's theory as able to accommodate Damasio's findings. However, the kind of decision-making to feature in the "somatic marker hypothesis" does not fit into the automatic and reflective binary set out by Korsgaard. This will be the primary focus of discussion in the remainder of this chapter. I will conduct an interdisciplinary analysis and integrate Korsgaard's ideas about reflection and reason with the work of Damasio and modern neuroscience. Through this process will emerge an integrative account of moral cognition.

4.2 Damasio, Korsgaard and a Spectrum of Deliberation

In the previous section, I summarised Damasio's "somatic marker hypothesis" which describes how feelings guide and constrain reasoning and motivate action (Damasio, 1996). I also discussed the implication of Damasio's work for sentimentalist views of motivation and briefly for Korsgaard's view of motivation. In this section I will continue to compare Korsgaard's and Damasio's accounts of decision-making. The focus will be on a key distinction between these accounts, namely that there is a solid binary between automatic and deliberative modes of acting. I will resolve this conflict

and show the compatibility of both accounts, which similarly involve a reflective process and a valuation process.

4.2.1 The Reflective Link Between Korsgaard and Damasio

The type of decision making described in the "somatic marker hypothesis" does not fit into Korsgaard's distinction between highly automatic and highly reflective modes of acting. Korsgaard's theory seems to involve a binary between reflective and non-reflective action. Korsgaard argues that either one is self-conscious and acts deliberatively or one is not and acts automatically. She says that "lower animals" are an example of the latter, where "its perceptions are its beliefs and its desires are its will" (Korsgaard, 1996, pp. 92-93). It is fixated on the environment, and its desires and passions "push it around", because it cannot reflect on the contents of its mind. One could argue that if a person refuses to reflect on their thoughts, and merely responds to their desires they are acting in this way. For example, if someone's emotional reaction (of say disgust) after witnessing an act of vandalism functions as their moral judgement, and they do not reflect any further, then they are acting automatically. Korsgaard argues however, that because humans can self-consciously reflect, they are necessarily faced with a normative question: "is this desire really a reason to act?" (Korsgaard, 1996, p. 93).

Damasio's work shows that the distinction between reflective and non-reflective action is not binary, but a matter of degree. According to the "somatic marker hypothesis" many of our judgements do employ reasons, but these reasons are grounded by an evaluation – i.e. a "feeling" – of what is important in a given situation (Damasio, 1994). This reasoning is more embodied and automatic, as opposed to being introspective and deliberative. However, both models of action involve a kind of reflection, in that they refer to a notion of identity. This reflection is either implicit or explicit, depending on how automatic or deliberate the involved judgement is.

There are reasons to think that Korsgaard is open to this idea. Different aspects of one's practical identity involve different cognitive demands. If someone conceives of themselves as a slave to their passions, they will endorse any and all their desires, and Korsgaard says the person will be "wanton" (Korsgaard, 1996, p. 101). Any reflective process will be as simple as identifying one's wants. On the other hand, as they conceive of themselves as a moral agent, required to take account of the reasons for

their actions and the perspectives of others, then reflection and deliberation will be more demanding. They will need to ask whether their reasons to act are universalizable and take the desires and well-being of others into account.

According to Damasio (1994), when someone is deciding, for example, between which deck to choose from in the Iowa Gambling task, or which charity to donate to, a somatic marker is generated which helps guide the process. This is automatic, one does not choose to experience this feeling, it merely arises because of the context one finds oneself. One does not reflect on and appeal to this feeling, it merely has its effect. Sometimes this feeling is so powerful that it forces one feels all but compelled to act in a certain way. Furthermore, somatic markers can affect actions unconsciously, meaning one can be moved without consciously reflecting on the motivating feeling (Damasio, 1994). There is still space for reflection and reasoning after one's initial reaction. Any subsequent reasoning process is automatically constrained and guided by feelings, but it is not inhibited.

As with Korsgaard's account, this feeling guided decision-making invokes a kind of reflection. This reflection however is immediate and partly unconscious, based on carefully moulded "habits of the heart" suiting one to the human ecosphere. Somatic markers become associated with particular actions and outcomes and reflect a sense of how the actions and outcomes affect the person. In this way, they enable a person to learn from previous experiences about how to act in accordance with who they are. When a context is reencountered, these somatic markers are reexperienced in order to avoid disadvantageous actions or encourage advantageous actions. The risky and futile actions of Damasio's vmPFC lesion patients indicate the inherent wisdom in "somatic markers", which they no longer feel (Damasio et al., 1990). When feelings are removed from the decision-making process, practical reasoning is impaired.

There is a second substantial link between Korsgaard's and Damasio's accounts of decision-making, centring on the importance of identity as a valuation of actions. Decision-making akin to the "somatic marker hypothesis" requires no explicit appeal to identity. However, as just indicated, the fact that one can learn from past experiences demonstrates a degree of personal continuity over time. A person can only learn from a situation, and adopt a pattern of responding to particular situations, if the earlier self is in some way continuous with ("identified with") the future self. This is the same

argument that Korsgaard uses in relation to reasons: just as using a reason entails a notion of personal identity, so too does forming a somatic marker.

Here we can see some key similarities with Damasio. For Korsgaard (1996, p. 101), what grounds or guides reasoning and decisions is an appeal to a valuation of oneself which is associated with actions. Similarly, according to the "somatic marker hypothesis", somatic markers are attached as positive or negative values to actions and outcomes (Damasio, 1994). Let us take an example, part of my practical identity is of a medical student, and this contributes to my sense of self-worth and directs my actions. During medical school students enter the hospital to begin their clinical training. It is essential in this training that students will have many challenging experiences, that will become embedded in themselves, and akin to Damasio, will direct their actions in future clinical scenarios. This may be learning how to talk to patients or their families when they have experienced immense tragedy, for example. Through this learning process a medical student will strive to become a good doctor by valuing certain actions over others. This process contributes to a medical student's conception of their doctorhood. If practical identity includes a description of what actions are "worth undertaking", then it seems that Korsgaard's account of morality is deeply similar to that of Damasio. The only issue that remains, is that for Damasio this is an automatic process, and for Korsgaard it is deliberate.

4.2.2 Automatic Judgements in the Spectrum of Deliberation

Decision-making akin to the "somatic marker hypothesis" does not fit into the binary distinction between automatic and reflectively deliberative judgements set by Korsgaard. However, these kinds of decision, despite being largely automatic maintain important properties laid out by Korsgaard. They are still reflective (albeit implicitly), and they still rely on a self-valuation of one's actions (which is the function of practical identities according to Korsgaard). Thus, it makes sense to revise Korsgaard's account of moral decision-making using the work of Damasio. This process indicates that perhaps the human ability to reflect and deliberate has many more gradations than first anticipated. In this section I will examine the neuroscience of automatic judgements, and I will relate these findings to the automatic-deliberative distinction between Damasio and Korsgaard's accounts of decision-making. Akin to Damasio's theory,

automatic judgements are more complicated than one would assume and integrate cognitive and emotive components.

Damasio is not the only neuroscientist to focus on automatic decision-making. There is significant overlap between Damasio's account of moral judgements and Greene's dual process theory. Greene argues that in more personal, emotional and temporally demanding circumstances, one relies on a kind of automatic judgement, that often produces seemingly deontological results (Greene, 2016). The vmPFC is the brain region implicated in Greene's account of these automatic judgements, as well as the critical link between deciding and somatic states in Damasio's hypothesis (Greene, 2016; Damasio, 1994). So, there is consistency across their works.

If one examines Greene's findings in the light of Damasio's hypothesis, one would argue that the vmPFC does not just represent automatic and emotional judgements but exists as a more complex intertwinement of emotion and reasoning. The philosopher James Woodward (2016) has offered such a reinterpretation of Greene's findings in the light of more recent neuroscientific studies. He argues that the vmPFC attributes subjective values to actions and judgements, in both utilitarian and deontological moral dilemmas (Woodward, 2016). And contrary to Greene's assertion that the vmPFC is primarily implicated in emotive and automatic processing, Woodward (2016) argues that instead it is essentially integrative in its treatment of cognition and emotion. He focuses on some of Greene's own results to critique his theory, specifically an fMRI study by Shenhav and Greene (2010). In this study the researchers examined the neural mechanisms of decision-making in response to complex moral dilemmas, which involved varying degrees of magnitude (the amount of lives lost) (Shenhav and Greene, 2010). They found that vmPFC activity correlated with the "expected moral values" of decision options, and hypothesised that in order to generate this valuation of moral action, the vmPFC must receive input from and modify affective representations (Shenhav and Greene, 2010). Woodward (2016) argues that these findings challenge the initial conception of Greene's dual process theory, that automatic judgements are essentially emotive, and instead show that the vmPFC plays both cognitive and emotive rolls.

A later fMRI study by Shenhav and Greene (2014), which was discussed in chapter one, appears to reinforce Woodward's reinterpretation. Participants were

presented sacrificial dilemmas while undergoing fMRI neuroimaging and were asked to choose between two responses, one was more utilitarian in nature and the other more deontological. Firstly, they were asked to make a utilitarian assessment, then an emotional assessment, and finally an "all things considered judgement", where they were asked: "Which do you find more morally acceptable?". The researchers found that activity in the vmPFC was greatest during the all things considered (i.e. integrative) moral judgement. Thus, its role was hypothesised to integrate and modulate affective signals into decision-making (Shenhav and Greene, 2014). Together these results indicate that the vmPFC carries out an integrated emotive and cognitive role. This coincides with Damasio's somatic marker hypothesis where the vmPFC computes subjective values, through an input and modulation of emotion. It also fits with Korsgaard's claim that moral judgement involves a conception of the self, considered as an "integrated whole".

Further evidence of the complex and integrative nature of automatic judgements is provided by a paper by Decety and Cacioppo (2012), that was also discussed in chapters one and two. EEG was used to study the temporal involvement of brain regions, while the participant made a basic moral judgement. Participants were exposed to two 3-frame videos, either depicting someone intentionally or unintentionally harming another, and were then asked to judge the intention of the actor. Using EEG, high-density event-related-potentials (ERP) were calculated. Increased activity in the posterior superior temporal sulcus (pSTS), the amygdala and the vmPFC was observed when participants watched the 'intentional harm' clip compared to the "unintentional harm" clip.[22] Importantly, these brain regions were activated in that order, with the pSTS first, then the amygdala and finally the vmPFC. The pSTS is involved in understanding the actions of others, including their intentions, representing a cognitive and categorisation process, which is necessary to proceed any judgement. The subsequent involvement of the amygdala represents an affective response to the categorised intention of the actor, and finally the involvement of the vmPFC, which is heavily connected to the amygdala represents an integration of this affective processing into a judgement. Decety and Cacioppo (2012) conclude that the amygdala acts a "gain-switch" to alert one about salient information and guide decision-making. This whole process occurs within 300ms, indicating that even in fast, automatic judgements, brain

[22] Refer to chapter one for summaries of these brain regions.

regions involved in reason and cognition both play important roles. These results are consistent with Damasio's hypothesis and give one a chronological understanding of how emotion influences moral decision-making.

These findings have implications for Korsgaard's theory. In accordance with Damasio's hypothesis, the evidence described here shows the importance of automatic judgements, and shows that we cannot regard automatic and reflective decisions as dichotomous. Although these kinds of judgements are not deliberative or reflective in the way set out by Korsgaard, they share essential aspects with her account. Decision-making akin to the "somatic marker hypothesis" is implicitly reflective, involving learning from past experiences, which corresponds with Korsgaard's assumption that humans are self-consciously reflective. It involves the intertwinement of reasoning and emotion, which constitutes a kind of feeling guided practical reasoning. This corresponds with Korsgaard's argument that passions, desires and internal argumentation can all be endorsed as potential reasons to act. Finally, Damasio's theory involves the valuation of ways of acting, which relates to Korsgaard's conception of practical identities. Although there are differences between Korsgaard and Damasio's accounts, they are compatible. And by applying some slight modifications to Korsgaard's theory, it can provide us with a neuroscientifically informed foundation for an integrative account of moral cognition. What has been examined so far are automatic moral judgements. There is, however, considerable evidence for more cognitively demanding and reflective judgements. In the next section, deliberative moral judgements will be examined regarding their relation to Damasio and Korsgaard and the spectrum of reflection.

4.2.3 Deliberative Judgements in the Spectrum of Deliberation

The kind of decision-making described by Damasio is automatic while that of Korsgaard is highly deliberative and reflective. It seems reasonable to suppose that there are kinds of judgements that sit functionally between the two. In these judgements cognitive elements would be important, but there would be less deliberate reflection. Imagine a scenario where you have to decide to donate a modest amount to a charity, but need to decide which charity to donate to. Suppose there are two donation boxes in front of you, one belongs to a national cancer society, and the other belongs to the local homeless shelter. In this case, one could act automatically and just donate to whichever

charity one feels moved toward at the time. Alternatively, one could make a more cognitively demanding judgement. That person would not need to stand back and carefully reflect and deliberate, but some basic reasoning could still factor into the decision. They may need to exert cognitive control over their emotions in order to make a judgement better supported by reasons. Perhaps initially, they are subconsciously drawn to the cancer society because it recently featured in an advertisement, but then they reason that their dollar would go further at the local homeless shelter. Emotion and cognition still play important roles here, but there is greater emphasis on subsequent deliberation over initial reactions.

Greene (2016) argues that when people give "utilitarian" responses to moral dilemmas they are thinking like this. Deliberation and cognitive control of more emotional brain regions play important roles in these decisions. Greene et al. (2004) implicates the dlPFC as one of these brain regions that works to resolve such a conflict. Somatic markers can still play a role in these decisions, however the emphasis rests on reasoning and executive control which occur after these feelings have had their effect. Damasio himself acknowledges that an extensive reasoning process can occur after somatic markers have their effect (Damasio, 1994). This process can be seen in a study by Suter and Hertwig (2011), who examined the relationship between deliberation time and the kind of judgement made by participants ("deontological" or "utilitarian"). Participants were randomly assigned to either a time-pressure condition where they were forced to give answers to moral dilemmas within 8 seconds, or a no-time-pressure condition, where they were given 3 minutes to deliberate. Participants under no time pressure were more likely to give consequentialist responses to the moral dilemmas compared to their time-pressured counterparts. Suter and Hertwig (2011) argued that under time pressure cognitive control mechanisms, which usually override or modify the effects of gut feelings, were lacking, leading to increased deontological decisions over consequentialist ones.

Suter and Hertwig's study helps to illustrate the differences between automatic decisions typical to the "somatic marker hypothesis" and more cognitively demanding decisions. Although regions such as the dlPFC play important roles, Greene's distinction between the emotive and cognitive is too extreme. The dual process model of moral judgement ignores the role emotions play in more reflective decision-making. Likewise, as we saw above when discussing the complex emotive and cognitive

functioning of the vmPFC, the dual process model underestimates the involvement of cognitive aspects in automatic judgements.

In more demanding contexts, Greene (2016) emphasises the role of cognitive brain regions over emotional ones. And the result is, according to Greene (2016), often a utilitarian decision. Woodward (2016) offers a valuable critique of this stringent categorisation by laying out a distinction between two different kinds of utilitarian judgements: "parametric" and "strategic". "Parametric" utilitarian decisions treat moral problems as having a "simple and transparent structure characterised by a few fixed and stable parameters" that are known for certain (Woodward, 2016). For example, take the overbridge variant of the trolley problem. A participant is told for certain that they are strong enough and easily able to push the man over the bridge, and that this will certainly stop the trolley. The dilemma is stripped of uncertainty, and comes down to what may be considered a simple maths problem: five lives versus one. Perhaps all that is happening "cognitively", is an exertion of control over one's emotional aversion to sacrificial killing. These are the kinds of dilemmas that Greene uses in his studies,[23] so it is not surprising that a limited set of cognitive brain regions are implicated during "utilitarian judgements" while emotional regions are deemphasised. "Strategic" utilitarian decisions on the other hand, are no less "utilitarian" in nature, but are far more practical and realistic. They involve deciding in the face of uncertainty, calculating probabilities, reasoning, attempting to retrieve more information from the unfolding scenario, problem solving, building an empathic relationship with potential victims and using empathy to detect any evil intentions (Woodward, 2016). In all these factors, as well as being more cognitively demanding in general, emotions and feelings can be a guide and can streamline thinking. In this respect, feelings are just as essential as cognition in practical reasoning. An example of "strategic" utilitarian decision-making, using the overbridge trolley dilemma, would involve a calculation or estimation of the uncertainty in the ability of the large man to stop the trolley. It could involve empathically relating to this man, asking him or thinking about the real-world impact of killing him. Does he have a family or people who rely on him? How about the five people tied up? Will the man even consider sacrificing himself willingly? This is far more complex than a simple maths problem, yet remains completely utilitarian.

[23] See for example: Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. and Cohen, J. D. (2004) 'The Neural Bases of Cognitive Conflict and Control in Moral Judgment', *Neuron*, 44(2), pp. 389-400.

Not only would cognitive brain regions be involved in these kinds of decisions, but emotional and empathy related brain regions would be involved as well (Woodward, 2016). Unfortunately, these are not the kinds of decisions focused on in the literature. However, it is likely that in the real world equivalents of such "utilitarian" or cognitively demanding moral decisions are "strategic" in nature and thus involve a far greater role for emotions than first anticipated.

As a final example, let us consider a study I outlined in chapter one: Feldmanhall et al.'s (2014) use of fMRI to examine which brain areas are involved in "difficult" moral dilemmas compared to "easy" moral dilemmas. To categorise dilemmas as either difficult or easy, participants were given a set of moral and non-moral dilemmas with two potential choices, and their responses were recorded. Moral dilemmas were distinguished from non-moral dilemmas, using the definition of moral cognition from Moll et al. (2008), that it must altruistically motivate behaviour. Dilemmas were regarded as difficult when there was little consensus on the "correct" answer (proportions picking one response in the binary choice were between 0.45-0.55) and at least 80% participants had to rate the dilemma as difficult (4 or 5, out of a 5 point-scale) (FeldmanHall et al., 2014). Dilemmas were regarded as easy if there was a consensus on the "correct" answer (either <0.20 or >0.80) and at least 80% of participants rated it as easy (1 or 2, out of 5). So, there was a clear and distinct difference between easy and difficult moral dilemmas.

fMRI data was collected from participants when they made judgements in response to easy moral and non-moral dilemmas, and difficult moral and non-moral dilemmas (FeldmanHall et al., 2014). When difficult moral dilemmas were compared to difficult non-moral dilemmas, the temporo-parietal junction (TPJ) was activated, and this activity extended into the temporal lobe to the temporal pole, whereas the vmPFC was relatively deactivated. The opposite was true when comparing easy moral decisions to easy non-moral decisions, with the vmPFC being relatively more activated, and the TPJ and the dlPFC being relatively more deactivated. Both of these results indicate specific differences between moral and non-moral thinking when difficulty is controlled for. When comparing difficult moral dilemmas to easy moral dilemmas, the bilateral TPJ and the right temporal pole were more activated in the difficult condition, and the vmPFC (and the left OFC) and the middle cingulate were relatively more activated in the easy condition. These results are consistent with Damasio's somatic

marker hypothesis, as the vmPFC is most sensitive to dilemmas where there is a low-cost and high-benefit option (i.e. one that is automatically obvious), which in turn can be encouraged through a feeling guided process (Damasio, 1994). On the other hand, it also shows that neural processing is different during more difficult and cognitively demanding scenarios, involving greater activity in the TPJ (FeldmanHall et al., 2014).

In chapter two I discussed a meta-analysis of 70 functional neuroimaging studies by Decety and Lamm (2007). In this meta-analysis TPJ has been implicated in four distinct domains, theory of mind (reasoning and attribution of the mental states of another, similar to cognitive empathy), empathy (defined in this paper as affective sharing with another), agency (explicitly or implicitly attributing ownership of actions to oneself or another) and attentional-shifting (shifting one's attention to a non-cued location or to detect and re-orientate to unexpected environmental changes). Hence this region is involved in many high-level cognitive processes important for social cognition and practical reasoning, which include emotional and cognitive components. The inclusion of attentional shifting here, corresponds to a reflective process. An actor is not narrowly focussed, but is assessing many salient stimuli, to factor in their decision. However, this process is not necessarily deliberative. This data supports the ideas put forward by Hughlings-Jackson about brain organisation, that higher-brain areas elaborate the functioning of simpler brain areas (Franz and Gillett, 2011).

With all this evidence in mind, responding to difficult moral dilemmas, involves a different network of brain regions from automatic and easy judgements, yet is not highly deliberative and reflective in the way Korsgaard's theory might seem to suggest. This kind of intermediate decision involves the intertwinement of emotion and reasoning, the collaboration of many different brain regions involved in many complex functions, and a modest level of reflection, represented by the shifting of attention between different salient factors in a dilemma. What this shows is that there are intermediate kinds of judgements in the spectrum of reflection.

Finally, there is one other kind of decision that sits on the far end of the reflective spectrum. It is the kind of highly deliberative and reflective decision-making, which one might associate with Korsgaard's account of morality (Korsgaard, 1996). This kind of highly reflective thinking is similar to when one does "arm-chair" philosophy. This is when one stands back from one's thoughts, chooses what to reflect

on and decides whether one's desires or certain arguments are truly reasons to act. This thinking is conscious and considered.

However, this kind of decision-making does not fit into the typical research paradigm of the neuroscience of moral cognition. In studies that use fMRI to examine brain activity when a participant makes a moral judgement, they are told to make a choice within a few seconds. This is due to technological constraints, with brain activity only from a specific and controlled section of time being recorded. Studies using EEG to study moral judgements, examine ERPs (event related potentials), which is a recording of brain activity for a fixed and controlled amount of time after a specific event (Decety and Cacioppo, 2012). This is repeated many times, to achieve an average recording of brain activity in response to that event. If one wants to study a highly reflective and deliberative judgement in response to a complex and deeply personal moral dilemma, one would be at a disadvantage using these neuroimaging techniques. In Chapter two I discussed the challenge of temporality, i.e. the problem of reducing moral judgements to short and basic responses to problems. This is a common challenge in the neuroscience of moral cognition. Consequently, what is often overlooked is a kind of common moral thinking that takes place over minutes, hours and even days. Hence, there is limited focus of this incredibly complex, extended and integrative processing in the moral cognition literature.

In Chapter three I raised a personal example about becoming a vegetarian, and I also asked you to think of your own personal examples of a detailed, interspersed and extended kind of moral thinking. Reflecting on these points it becomes obvious, that despite receiving a limited focus in the literature, highly reflective decision-making plays important roles in our lives. Also, in chapter three I introduced a study by Paxton et al. (2012) that examined the role of reflection in altering one's moral judgements and responses to reasoning. It was found that when more time was given to think about a controversial moral scenario (allowing for greater deliberation and reflection) only then did argument strength effect the permissibility rating of the scenario given by participants. This shows that reflection and deliberation do occur, especially when significant time is given to contemplate moral problems, and that they do affect how one reasons and how one uses arguments to inform their judgements.

4.2.4 Modifying Korsgaard's Theory According to a Reflective Spectrum of Deliberation

In this chapter I have described three kinds of decisions. Firstly, automatic, feeling guided decisions, best described in Damasio's "somatic marker hypothesis". Secondly, more cognitively demanding, and basically reflective decision-making, at an intermediate level. And finally, the highly deliberative and consciously reflective decision-making described by Korsgaard in her account of moral decision-making. These three processes all have different levels of automaticity and deliberation yet maintain the essential aspects of Korsgaard's moral philosophy. They are all reflective in some way yet vary depending on how implicit or explicit this reflection is. Secondly, they all involve emotion and cognition in an integrative way.

So, in response to these findings I propose a modification of Korsgaards theory. Before laying out this framework, let us briefly review her theory. Korsgaard's moral philosophy is a proposed solution to the normative question. She argues that normativity arises as we reflect on our thoughts and endorse particular reasons to act (Korsgaard, 1996). The reflective structure of self-consciousness, she maintains, forces one to form a conception of oneself, and this self-conception governs which reasons are endorsed. In her theory, if a person does not reflect and endorse reasons in this way then they are not truly "acting"; they are merely being pushed around by their desires.

As a modification of her theory, instead of there being a stringent dichotomy between highly automatic and highly deliberative ways of acting, there is a spectrum of reflection. On one end of the spectrum are automatic and implicitly reflective decisions, and on the other end are deliberate and explicitly reflective decisions. There is significant evidence illustrating the gradations of possible ways humans can engage with ethical issues, which range through this spectrum. As an example of such an automatic decision, one might find oneself in a position where one decides to help a lost stranger in a large city. This decision to help is likely automatic and intuitive, existing merely as a feeling guided response. However, as we saw when discussing Damasio, these feelings and intuitions arise from an implicitly reflective process, specifically from automatically learning from past experiences and building these into the basic ways one responds to the world (Damasio, 1994). On the other, deliberative end of the spectrum, one might decide who to vote for in a democratic election, which

would obviously have many widespread ethical consequences. This decision may involve a great deal of reflection and deliberation, including an engagement with one's own practical identity to govern this decision.

Importantly, both processes are reflective in some way, and thus relate to the first of Korsgaard's two major psychological assumptions. Secondly, both processes involve a valuation of actions. The former does this automatically, by attaching positive or negative signals to certain ways of acting based on past experiences. This learning process underlies how to be a good citizen in a society, or a good member of one's profession, or a good partner or parent. Hence, certain ways of acting become valued over others and embedded in the actor. The deliberative way of deciding, does this by reflecting on reasons to act and endorsing them with reference to one's practical identity, i.e. a valuation of oneself, the kind of life one finds worth living and the actions worth undertaking (Korsgaard, 1996, p. 101). So, when Korsgaard argues that normativity arises as we reflect on our actions and the possibilities available and endorse certain reasons, this process itself is complex. It can happen on automatic, intermediate and highly deliberative levels.

The third important feature to note is that at each level of this reflective and valuation-based decision making, feelings and cognition, or emotions and reasoning play important and intertwined roles. Through an examination of the work of Damasio, Korsgaard and other studies in cognitive neuroscience, it appears that the dichotomy between emotion and reason is a false one. Damasio (1994) argues that feelings as much as cognition are deeply involved in automatic and everyday practical reasoning. Emotive and empathic brain regions, similarly, are involved in more cognitively demanding decisions that warrant deliberation (FeldmanHall et al., 2014). And finally, Korsgaard argues that during rational reflection, emotions, desires, feelings and passions can all be endorsed as potential reasons to act. So, it appears that moral cognition is integrative at all cognitive levels in terms of emotions and reasoning.

In this chapter I have proposed a modification of Korsgaard's moral philosophy, using the work of Damasio and key neuroscience studies introduced in chapter one. The aim was to carry out an exercise in the interdisciplinary analysis indicated in the first three chapters. In the first three chapters I used moral philosophy to critically analyse neuroscience, whereas the focus of this chapter was to reciprocally critique theory in

moral philosophy using neuroscience. In this analysis I developed an integrative model of moral reasoning that shows how both sentiments and reasons are integrated in our moral judgements. This account abides by the three features I refer to as an “integrative account of morality”.

Conclusion

The objective of this thesis is to discuss the need for an account of moral cognition that is integrative in character and interdisciplinary in approach. In the first three chapters I critically analysed the neuroscience of moral cognition using moral philosophy. In the fourth chapter I undertook an interdisciplinary analysis, including a reciprocal critique of moral philosophy using neuroscience.

In chapter one I reviewed recent neuroscientific evidence concerning moral cognition. The literature shows that moral cognition involves a coordination of brain processes, which operate as a physically integrated network. However, many of these findings can be interpreted flexibly, and can be used to argue for multiple models of moral cognition. Such models differ in the proposed roles of reason and emotion and in the proposed organisation of brain networks. To aid in the interpretation of results neuroscience needs to turn to the world of moral theory.

In the second chapter I discussed the pervasive link between theories in moral philosophy and models of moral cognition. Moral philosophy is necessary as a notion of "morality" is required prior to any neuroscientific study, in order to give the study a clear focus. Many researchers in this field are aware of this, and often explicitly take up philosophical presuppositions. However, they do so in cursory ways. For example, there is an overreliance on traditional and simplified versions of moral theories in the neuroscience literature. Furthermore, current neuroscientific techniques are unable to describe the temporal and holistic nature of moral cognition. Any analysis which relies solely on these techniques is missing normal aspects of moral thinking. Therefore, neuroscience can only be part of an interdisciplinary approach. A theoretical neuroscience informed by a nuanced moral philosophy is needed to start to approach a description of the relationship between the brain and morality. These challenges facing the neuroscientific study of morality indicate the need for an integrative and interdisciplinary account.

In chapter three, I showed how the relationship between the brain and morality can be broken down into two perspectives. An explanatory account attempts to describe the causally related brain processes behind moral thinking, from a scientific perspective. Whereas a normative account of morality attempts to explain the inherent

obligatory nature of moral claims, and this is from first-person perspective. The moral philosophy of Christine Korsgaard was discussed as it is particular effective in introducing the first-person or normative account of morality. I showed the insufficiency of explanatory accounts of morality in addressing the normativity behind moral claims, hence the need for moral philosophy and an interdisciplinary analysis.

Finally, this "integrative" work culminated in chapter four where I underwent such an interdisciplinary analysis. Specifically, I discussed how Korsgaard's understanding of rational reflection is both supported and moderated by the work of contemporary neuroscience, particularly the work of the neurologist Antonio Damasio. The model that was proposed moves towards a more accurate account of moral reasoning and contributes to an understanding of the relationship between the brain and morality.

In the introduction I summarised the three features of an integrative account of moral cognition. All three of these features represent a way the framework is "integrative". Let us again review what they are:

1. Neuroscientifically integrated: Moral Cognition implicates a network of brain regions. These regions operate with integrative brain processes and as such should be described using theories of higher-level cognition. The intertwinement of reason and emotion is essential in this framework. This represents a rejection of a strong distinction between these phenomena in this context. This feature was indicated in chapter one, where moral cognition was shown to operate as a complex network, and in chapter two, through a discussion of the inadequacy of neuroscience in describing the temporal extent and holistic nature of moral cognition.
2. Integrated across disciplines (as in 'inter-disciplinary'): From a methodological standpoint any comprehensive description of moral cognition ought to refer to moral philosophy, neuroscience and psychology (at least) and integrate their respective methods, generating a synthesis of perspectives. This feature was also indicated in chapter two, as a synthesis in perspectives can overcome challenges in the conceptualisation, testing and interpretation of any findings from empirical neuroscience.

3. Integrated explanatory and normative accounts of morality: To form a complete account of our moral lives, both explanatory and normative descriptions are necessary. Any explanation of moral cognition would be amiss if it did not account for the normative appearance of moral claims on us. This feature was indicated in chapter three, with a discussion of Korsgaard's neo-Kantian account of moral obligation.

Respecting these features, I have set out a framework of an integrative account of moral cognition. The framework is based on an integration of Korsgaard's ideas about reason and reflection with current empirical and theoretical neuroscience. According to this framework, moral cognition can vary depending on the level of automaticity and deliberation yet remains essentially reflective across this spectrum. This can range from an implicitly reflective process, where past experiences are incorporated into one's intuitive ways of acting using feelings that guide practical reasoning and acting, to an explicitly reflective process, where one can consciously deliberate on one's thoughts, feelings and internal argumentation. What governs which thoughts are endorsed as reasons to act, or which actions are encouraged or discouraged via feelings, depends on a valuation process. At the automatic level, certain actions and judgements are valued or disvalued based on past experiences representing an intuitive understanding of correct ways to act in specific contexts. At the deliberative level, this involves reference to one's practical identity, which contains a valuation of which actions ought to be carried out in certain contexts, including the actor's role within them.

Furthermore, at each point along this spectrum of reflection, reason, feeling and social cognition are intertwined and operate together to carry out the "higher-function" of moral cognition. Drawing from Hughlings-Jackson's ideas about brain organisation, basic sensorimotor processes are represented and meta-represented at varying levels, integrating with other brain regions, to achieve this (Gillett and Franz, 2014). The view that moral cognition functions as a network, with feelings and reasoning being inseparable and collaborative elements of practical reasoning, dispels any stringent dichotomy between reason and emotion. This basic framework was generated through a synthesis of perspectives, using neuroscience, philosophy and psychology in accordance with the imperative of being methodologically interdisciplinary.

Not only does this framework, on a basic level, produce an explanatory account of morality, it remains open to a normative account of morality. The account of moral cognition present is essentially reflective and relies on a valuation of actions to guide judgements. This is true in more automatic and deliberative judgements. These two elements are represented in Korsgaard's two psychological assumptions at the core of her theory. Humans are self-consciously reflective beings, and this forces us to examine our thoughts and ask ourselves whether they are good reasons to act. It also forces us to form a conception of ourselves, which Korsgaard (1996, p. 101) calls a "practical identity", a "description under which you find your life to be worth living and your actions to be worth undertaking". Normativity, the sense of "ought" underlying moral claims on us, arises when we reflect in the contents of our minds, and using our practical identity, ask whether the thought constitutes a good reason to act. If we endorse the reason, we are self-legislating, and we now have an obligation to act in that way. In this integrative account of moral cognition, there is a focus on both normative and explanatory components.

How is this framework similar and different from previously existing models of moral cognition? It contains many similarities with Greene's "dual process model of moral cognition" but remains essentially different. The dual process model includes an account of automatic and deliberative judgements and is multifaceted, involving important roles for both reasoning and emotions (Greene, 2016), which are points of agreement with the framework proposed in this thesis. However, in the dual process model, there is a hard distinction between automatic and deliberative judgements, and between emotion and reasoning, which implies that these faculties are disintegrated. The framework proposed by this thesis argues instead of the essential integration of brain processes, including processes involving feelings, reasoning and social cognition in moral judgements. For the same reason, it is opposed to the models of moral cognition formulated by Kohlberg, a rationalist and Prinz, a sentimentalist. These theorists argue for the primacy of either reason and deliberation or sentiments and emotions respectively in moral judgements (Kohlberg, 1971; Prinz, 2007). While there some is agreement with both cases, in that reasoning and emotions are constituent (and intertwined) parts in moral judgements, the rejection of Kohlberg and Prinz's models' rests in the complete dismissal of one of these faculties in favour of the other.

Finally, the framework of an integrated moral cognition has the most in common with the models of moral cognition put forward by Casebeer and Churchland (2003), and Moll et al (2008). Casebeer and Churchland (2003) argue in favour of a model of moral cognition best represented by a neo-Aristotelian virtue theory, where morality is concerned with what people should think and do to "function well as human beings".[24] They describe their model of moral cognition as "a large-scale brain affair depending on the appropriate coordination of many areas", which, they claim is better represented by a neo-Aristotelian virtue theory, than by a form of rationalism, sentimentalism or utilitarianism (Casebeer and Churchland, 2003). Their description of a multifaceted model, involving reason, emotion and social cognition has much in common with the framework presented in this thesis, with its focus on integrational brain processes. The model of moral cognition proposed by Moll et al. (2008), is still primarily sentimentalist as it focuses on moral motivators but is more nuanced and has more in common with the framework proposed in this thesis. Moll et al. (2008) proposes that 'emotion and cognition are nondissociable elements underlying moral motivations, and that such motivations are represented within cortico-limbic neural assemblies'. This view maintains that both emotion and reason are important in moral cognition, and that they are integrated in this function.

Obviously, what has been developed is only a broad and basic framework for understanding moral cognition. Much more work can be done in all applicable fields. In chapter three I clarified the overarching aim of this model, which is to overcome the challenges that face the neuroscientific study of moral cognition. This analysis will aid in the development of a more robust and pluralistic model, that reveals a critical link between reason and reflection. This is not intended to be a comprehensive or exclusive model of moral cognition. This work would likely be further enhanced with the addition of more perspectives in moral philosophy, and also a more specific focus on aspects of the moral network and how they operate together specifically. One such example for future study could involve analysing the role of empathy and human developmental ecology from the perspective of an interdisciplinary and integrative account of moral cognition.

[24] For another theory of moral cognition from an Aristotelian perspective see Gillett, G. (2018) *From Aristotle to cognitive neuroscience.* Cham, Switzerland: Palgrave Macmillan.

The study of moral cognition is important as knowledge about the neuroscience underlying morality will inform and guide practical ethics. For example, neuroscientific and psychological information about how humans' reason, intuit and reflect upon moral dilemmas can reveal psychological pitfalls that must be avoided if we want to be ethically consistent and robust. Some ethicists even go so far as advocating for biological enhancement to improve our moral decision-making (Persson and Savulescu, 2008), and an understanding of moral cognition is essential here.

However, if neuroscience is to further a study of moral cognition it needs to recognise the shortcomings and limits of the presuppositions made about the nature of morality, some of which come out in experimental restraints and others through a cursory use of moral philosophy. It is worthwhile for the neuroscience to be critiqued and developed by moral philosophy in this way, so that it performs its descriptive role as best as possible. Furthermore, a philosophically informed neuroscience will set reasonable limits to any normative account of morality.

When I talk about the need for an interdisciplinary approach such as this, which integrates data and theory, empirical neuroscience and moral philosophy, I think back towards Kant's aspiration for the study of ethics. He thought that morality required a systematic and methodical study, not unlike science. In the introduction I began with an excerpt from the end of Critique of Practical Reason where Kant argues for the synergism of the empirical and the rational (Kant, 1993). In the context of moral cognition, this passage empathises a need for a collaboration of empirical neuroscience with theoretical neuroscience and moral philosophy:

> We have at hand examples of the morally judging reason. We may analyse them into their elementary concepts, adopting, in default of mathematics, a process similar to that of chemistry, i.e., we may, in repeated experiments on common sense, separate the empirical from the rational, exhibit each of them in a pure state, and show what each by itself can accomplish. Thus we shall avoid the error of a crude and unpracticed judgment and (something far more important) the extravagances of genius, by which, as by the adepts of the philosopher's stone, visionary treasures are promised and real treasures are squandered for lack of methodical study and knowledge of nature. In a word, science (critically sought and methodically directed) is the narrow gate that leads to the doctrine of wisdom, when by this is

> understood not merely what one ought to do but what should serve as a guide to teachers in laying out plainly and well the path to wisdom which everyone should follow, and in keeping others from going astray. It is a science of which philosophy must always remain the guardian...
>
> (Kant, 1993, pp. 170-171)

This thesis stands as a call for such a collaborative perspective on moral cognition. Every now and then a field, like that of moral cognition, with deep roots in multiple subjects needs a breath. It requires a moment of reflection of its goals and the methods to achieve them. This reflection informs any needed recalibrations. My hope is that an integrative account of moral cognition is such a reflection, where the neuroscientifically integrative, the interdisciplinary and the normative aspects of studying human morality are reemphasised. What now remains is for the momentous empirical and theoretical work to be continued, with greater self-awareness and frequent instances of reflection.

References

Bechara, A., Damasio, A. R., Damasio, H. and Anderson, S. W. (1994) 'Insensitivity to future consequences following damage to human prefrontal cortex', *Cognition,* 50(1-3), pp. 7-15.

Bechara, A., Tranel, D., Damasio, H. and Damasio, A. R. (1996) 'Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex', *Cerebral cortex,* 6(2), pp. 215-225.

Bloom, P. (2016) *Against empathy: The case for rational compassion.* The Bodley Head.

Casebeer, W. D. and Churchland, P. S. (2003) 'The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making', *Biology and philosophy,* 18(1), pp. 169-194.

Damasio, A. (1994) *Descartes' error: emotion, reason, and the human brain.* G. P. Putnam's Sons.

Damasio, A. R. (1996) 'The somatic marker hypothesis and the possible functions of the prefrontal cortex', *Phil. Trans. R. Soc. Lond. B,* 351(1346), pp. 1413-1420.

Damasio, A. R., Tranel, D. and Damasio, H. (1990) 'Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli', *Behavioural Brain Research,* 41(2), pp. 81-94.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M. and Damasio, A. R. (1994) 'The return of Phineas Gage: clues about the brain from the skull of a famous patient', *Science,* 264(5162), pp. 1102-1105.

Davidson, R. J., Putnam, K. M. and Larson, C. L. (2000) 'Dysfunction in the neural circuitry of emotion regulation--a possible prelude to violence', *Science,* 289(5479), pp. 591-594.

Decety, J. and Cacioppo, S. (2012) 'The speed of morality: a high-density electrical neuroimaging study', *Journal of Neurophysiology,* 108(11), pp. 3068-3072.

Decety, J. and Lamm, C. (2007) 'The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition', *The Neuroscientist,* 13(6), pp. 580-593.

Decety, J. and Yoder, K. J. (2016) 'Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others', *Social Neuroscience,* 11(1), pp. 1-14.

FeldmanHall, O., Mobbs, D. and Dalgleish, T. (2014) 'Deconstructing the brain's moral network: Dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex', *Social Cognitive and Affective Neuroscience,* 9(3), pp. 297-306.

Foot, P. 1978. Virtues and Vices and Others Essays in Moral Philosophy. Berkeley: University of California Press.

Franz, E. A. and Gillett, G. (2011) 'John Hughlings Jackson's evolutionary neurology: A unifying framework for cognitive neuroscience', *Brain,* 134(10), pp. 3114-3120.

Fumagalli, M. and Priori, A. (2012) 'Functional and clinical neuroanatomy of morality', *Brain: A Journal of Neurology,* 135(7), pp. 2006-2021.

Gillett, G. (2018) *From Aristotle to cognitive neuroscience.* Cham, Switzerland: Palgrave Macmillan.

Gillett, G. and Franz, E. (2014) 'Evolutionary neurology, responsive equilibrium, and the moral brain', *Consciousness and Cognition: An International Journal*, 45, pp. 245-250.
Greene, J. (2008) 'The Secret Joke of Kant's Soul', in Sinnott-Armstrong, W. (ed.) *Moral Psychology*: MIT Press, pp. 35-80.
Greene, J. D. (2014) *Moral tribes: Emotion, reason, and the gap between us and them.* Penguin.
Greene, J. D. (2015) 'The rise of moral cognition', *Cognition,* 135, pp. 39-42.
Greene, J. D. (2016) 'Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics', in Liao, S.M. (ed.) *Moral Brains: The Neuroscience of Morality*. New York: Oxford University Press, pp. 119-149.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. and Cohen, J. D. (2004) 'The Neural Bases of Cognitive Conflict and Control in Moral Judgment', *Neuron,* 44(2), pp. 389-400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. and Cohen, J. D. (2001) 'An fMRI investigation of emotional engagement in moral judgment', *Science,* 293(5537), pp. 2105-2108.
Haidt, J. (2001) 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment', *Psychological Review,* 108(4), pp. 814-834.
Hampton, R. R. (2001) 'Rhesus monkeys know when they remember', *Proceedings of the National Academy of Sciences,* 98(9), pp. 5359-5362.
Harlow, J. M. (1993) 'Recovery from the passage of an iron bar through the head', *History of Psychiatry,* 4(14), pp. 274-281.
Harlow, J. M. (1999) 'Passage of an iron rod through the head', *The Journal of Neuropsychiatry and Clinical Neurosciences,* 11(2), pp. 281-283.
Hume, D. (2007) *A treatise of human nature: A critical edition.* Oxford University Press.
Kahane, G. (2015) 'Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment', *Social neuroscience,* 10(5), pp. 551-560.
Kant, I. (1993) *Critique of practical reason.* 3rd edn. Upper Saddle River, New Jersey: Prentice-Hall.
Kant, I. (2002) *Groundwork for the Metaphysics of Morals.* Yale University Press.
Kennedy, D. P. and Adolphs, R. (2012) 'The social brain in psychiatric and neurological disorders', *Trends in Cognitive Sciences,* 16(11), pp. 559-572.
Koenigs, M., Kruepke, M., Zeier, J. and Newman, J. P. (2011) 'Utilitarian moral judgment in psychopathy', *Social cognitive and affective neuroscience,* 7(6), pp. 708-714.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. and Damasio, A. (2007) 'Damage to the prefrontal cortex increases utilitarian moral judgements', *Nature,* 446(7138), pp. 908.
Kohlberg, L. (1971) 'From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development', *Cognitive development and epistemology*: Elsevier, pp. 151-235.
Korsgaard, C. (1989) 'Personal identity and the unity of agency: A Kantian response to Parfit', *Philosophy and Public Affairs,* 18(2), pp. 101-132.
Korsgaard, C. (1996) *The sources of normativity.* Cambridge University Press.
Korsgaard, C. M. (2009) *Self-constitution: Agency, identity, and integrity.* Oxford University Press.

Lane, R. D., Chua, P. M. L. and Dolan, R. J. (1999) 'Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures', *Neuropsychologia,* 37(9), pp. 989-997.
Lind, G. 2008. The meaning and measurement of moral judgment competence: A dual-aspect model.
Mackie, J. L. (1990) *Ethics: Inventing right and wrong.* Penguin UK.
Macmillan, M. and Lena, M. L. (2010) 'Rehabilitating Phineas Gage', *Neuropsychological rehabilitation,* 20(5), pp. 641-658.
Mendez, M. F. (2009) 'The neurobiology of moral behavior: Review and neuropsychiatric implications', *CNS Spectrums,* 14(11), pp. 608-620.
Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J. n., Andreiuolo, P. A. and Pessoa, L. (2002) 'The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions', *Journal of neuroscience,* 22(7), pp. 2730-2736.
Moll, J., De Oliveira-Souza, R. and Zahn, R. (2008) 'The neural basis of moral cognition', *Annals of the New York Academy of Sciences,* 1124(1), pp. 161-180.
Mulert, C., Jäger, L., Schmitt, R., Bussfeld, P., Pogarell, O., Möller, H.-J., Juckel, G. and Hegerl, U. (2004) 'Integration of fMRI and simultaneous EEG: towards a comprehensive understanding of localization and time-course of brain activity in target detection', *NeuroImage,* 22(1), pp. 83-94.
Olatunji, B. O., Puncochar, B. D. and Cox, R. (2016) 'Effects of experienced disgust on morally-relevant judgments', *PLoS ONE Vol 11(8), 2016, ArtID e0160357,* 11(8).
Padoa-Schioppa, C. and Cai, X. (2011) 'The orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives', *Annals of the New York Academy of Sciences,* 1239(1), pp. 130-137.
Parfit, D. (1984) *Reasons and persons.* Oxford University Press.
Paxton, J. M., Ungar, L. and Greene, J. D. (2012) 'Reflection and reasoning in moral judgment', *Cognitive Science,* 36(1), pp. 163-177.
Persson, I. and Savulescu, J. (2008) 'The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity', *Journal of Applied Philosophy,* 25(3), pp. 162-177.
Pizarro, D. A. and Bloom, P. (2003) 'The intelligence of the moral intuitions: A comment on Haidt (2001)', *Psychological Review,* 110(1), pp. 193-196.
Prehn, K., Wartenburger, I., Mériau, K., Scheibe, C., Goodenough, O. R., Villringer, A., van der Meer, E. and Heekeren, H. R. (2007) 'Individual differences in moral judgment competence influence neural correlates of socio-normative judgments', *Social cognitive and affective neuroscience,* 3(1), pp. 33-46.
Prinz, J. (2007) *The emotional construction of morals.* Oxford University Press.
Prinz, J. (2011) 'Against empathy', *The Southern Journal of Philosophy,* 49(s1), pp. 214-233.
Prinz, J. (2016) 'Sentimentalism and the moral brain', in Liao, S.M. (ed.) *Moral brains: The neuroscience of morality*. New York, NY: Oxford University Press; US, pp. 45-73.
Saxe, R. and Wexler, A. (2005) 'Making sense of another mind: the role of the right temporo-parietal junction', *Neuropsychologia,* 43(10), pp. 1391-1399.
Shenhav, A. and Greene, J. D. (2010) 'Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude', *Neuron,* 67(4), pp. 667-677.

Shenhav, A. and Greene, J. D. (2014) 'Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex', *The Journal of Neuroscience,* 34(13), pp. 4741-4749.
Singer, P. (1997) 'The drowning child and the expanding circle', *New Internationalist,* (no. 289).
Singer, T. (2006) 'The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research', *Neuroscience & Biobehavioral Reviews,* 30(6), pp. 855-863.
Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J. and Frith, C. D. (2004) 'Empathy for pain involves the affective but not sensory components of pain', *Science,* 303(5661), pp. 1157-1162.
Slote, M. (2007) *The ethics of care and empathy.* Routledge.
Sowell, E. R., Thompson, P. M., Tessner, K. D. and Toga, A. W. (2001) 'Mapping continued brain growth and gray matter density reduction in dorsal frontal cortex: Inverse relationships during postadolescent brain maturation', *Journal of Neuroscience,* 21(22), pp. 8819-8829.
Steinbeis, N., Haushofer, J., Fehr, E. and Singer, T. (2016) 'Development of Behavioral Control and Associated vmPFC-DLPFC Connectivity Explains Children's Increased Resistance to Temptation in Intertemporal Choice', *Cerebral Cortex,* 26(1), pp. 32-42.
Suter, R. S. and Hertwig, R. (2011) 'Time and moral judgment', *Cognition,* 119(3), pp. 454-458.
Thomson, J. J. (1985) 'The Trolley Problem', *The Yale Law Journal,* 94(6), pp. 1395-1415.
Wagner, N. F., Chaves, P. and Wolff, A. (2017) 'Discovering the Neural Nature of Moral Cognition? Empirical, Theoretical, and Practical Challenges in Bioethical Research with Electroencephalography (EEG)', *Journal of Bioethical Inquiry,* 14(2), pp. 299-313.
Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V. and Rizzolatti, G. (2003) 'Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust', *Neuron,* 40(3), pp. 655-664.
Woodward, J. (2016) 'Emotion versus cognition in moral decision-making', in Liao, S.M. (ed.) *Moral brains: the neuroscience of morality*. New York: Oxford University Press, pp. 87-116.
Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A. and Saxe, R. (2010) 'Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments', *Proceedings of the National Academy of Sciences,* 107(15), pp. 6753-6758.
Zheng, H., Lu, X. and Huang, D. (2018) 'tDCS over DLPFC leads to less utilitarian response in moral-personal judgment', *Frontiers in Neuroscience,* 12(MAR).