

**Do we have a unified moral faculty? No.**

Yuhan Fu

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Faculty of Arts and Humanities

Department of Philosophy

Submission Date: August, 2023

# Acknowledgements

Acknowledgements should be thesis long. There are too many people whom I will be forever indebted. It would be morally wrong not to thank them properly. But here we go.

I am blessed to do my PhD at the dearest Sheffield philosophy department. First thing first, the most importantly, shout out to the best supervisors in the world: Ryan Byerly, Gerardo Viera and Stephen Laurence. Their insightful comments and feedback have helped me with shaping my research, and made me a better philosopher. I have also greatly benefited from Shaun Nichols’ visit to Sheffield during this spring semester. Can’t forget the legend: the former Research Support Officer Patrizia Baldi, whose supports over the past few years has been immeasuarable.

To my philosofriends at King’s and Sheffield: Sam Heffron, Katie Sparks, Leonard Weiss, Patricia Schöllhorn-Gaar, Kim Engel, Kimi Chen, Yifan Mei, Jimmy Lenman, Jules Holroyd, Yonatan Shemmer, Luca Barlassina, Lewis Brooks, Miklos Kurthy, Alex Duval, Viktoriia Kononova, Jack Herbert, Anna Klieber, William Morgan, Charlie Potter, James Brown, Edward Armitage, James Turner, Henry Roe, George Surtees, Ben Jenkins, and Samuel Chan. All the proofreading, arguments, chats, companies, coffee, beers, and cuddles have made my challenging MA and PhD philosophy journey so uplifting and fun. I would not finish this thesis without all the philosophical debates and infinite supports.

Last but not the least, I would like to thank my brother Shengkun Fu who has been my biggest fan since his birth. To my mum Xiaoying Deng and my dad Shibo Fu: I know it’s hard to digest my impulsive decision of doing a philosophy PhD in another side of the planet, but you support me regardless, this thesis is for you.

# 

# Abstract

This thesis is going to scrutinise what current cognitive science can tell us about our moral cognition, to explore what psychological mechanism(s) facilitate(s) moral cognition, and to investigate the question of whether we have an innate, unified, and domain-specific psychological mechanism for moral cognition. I will argue that current evidence suggests that our moral cognition is a result of our general complex and flexible learning processes, which also perform other psychological tasks. The thesis will be divided into three parts. The first part consists of three chapters, which will be on the role of emotions in moral cognition: chapter 1 will focus on Jesse Prinz’ constitution model account, chapter 2 will focus on Shaun Nichols’ sentimental rules’ account, chapter 3 will be a general discussion on the role of emotions in moral judgements. The second part focuses on two theories that scaffold morality on the ground of evolutionary psychology: chapter 4 is on Chomskyan moral nativism and chapter 5 will be on cooperation-based mutualistic account of moral theory. I will argue that evolutionary psychology cannot indicate that human moral cognition is domain-specific. In my last part, I will turn to discuss the relationship between general learning capacity and moral cognition. In chapter 6, I will argue that although Shaun Nichols is right that moral cognition is not innate, we are not Bayesian learners as he argues. Chapter 7 will be my last chapter, in which I will explore whether machines can engage with moral cognition. I will argue that based on current performance of large language models, machines are able to engage with moral cognition, which further refutes the domain-specificity of moral cognition. I will summarise what I have argued in my conclusion.

# Declaration

*I, Yuhan Fu, confirm that the Thesis is my own work. I am aware of the University’s Guidance on the Use of Unfair Means (*[*www.sheffield.ac.uk/ssid/unfair-means*](http://www.sheffield.ac.uk/ssid/unfair-means)*). This work has not been previously been presented for an award at this, or any other, university.*

Table of Contents

[Acknowledgements 2](#_Toc141819189)

[Abstract 3](#_Toc141819190)

[Declaration 4](#_Toc141819191)

[Introduction 6](#_Toc141819192)

[Part 1: Moral Sentimentalism and the Role of Emotion in Moral Judgement 18](#_Toc141819193)

[Chapter 1 Worries for Constitution Sentimentalists’ Account of Moral Judgement 20](#_Toc141819194)

[Chapter 2 Mind the Gap: Why Empirical Studies Do Not Support the Sentimental Rules Account 37](#_Toc141819195)

[Chapter 3 Wrap Up: The Role of Emotions in Moral Judgements 60](#_Toc141819196)

[Part 2: Moral Competence and Moral Learning 70](#_Toc141819197)

[Chapter 4 Do We Have an Innate Universal Moral Grammar 72](#_Toc141819198)

[Chapter 5 Cooperation and Morality: Cooperation is Not the Whole Story of Moral Cognition 96](#_Toc141819199)

[Part 3: General Learning Mechanism, Machine and Morality 110](#_Toc141819200)

[Chapter 6 Problems with Moral Bayesianism 112](#_Toc141819201)

[Chapter 7 The Emotionless Machine and Its Normative Tail: Can AI Engage in Moral Cognitions? 139](#_Toc141819202)

[Conclusion 160](#_Toc141819203)

[References 162](#_Toc141819204)

Do we have a unified moral faculty?

No.

# Introduction

“I’ve done a lot of bad stuff through the years. Now I’m paying the price but there’s so many things I’ll never get a chance to do: smoke a cigarette, use a fake I.D., shave a swear word in my hair…”

― Bart Simpson

Bart is just one of us: we all have a good understanding of what kind of actions are bad (smoke a cigarette, use a fake I.D., shave a swear word in your hair for example). We human beings possess a very impressive capacity which is known as moral cognition: we are able to detect moral norms from our environment since we are young, internalise these moral norms and make moral judgements of our own and others’ actions based on them. Sometimes we agree with each other, and take our moral judgements seriously, but we sometimes also fail to agree with each other and fail to take these moral norms seriously as well.

This thesis is going to scrutinise what current cognitive science can tell us about our moral cognition, to explore what psychological mechanism(s) facilitate(s) moral cognition, and to investigate the question of whether we have an innate, unified, and domain-specific psychological mechanism for moral cognition. The ultimate goal of this thesis is to convince my readers that although we demonstrate moral capacity, there isn’t, at least in the sense that there is no innate psychological system that *only* accounts for moral cognition. Instead, I will argue that current evidence suggests that our moral cognition is a result of our general complex and flexible learning processes, which also perform other psychological tasks. In other words, I argue that our moral faculty is domain-general, informationally integrated, and cognitively penetrable. Hence, in my thesis, I will reject claims which try to provide a unified and domain-specific account for moral cognition.

1. Clarificatory work: central question, methodology, and relevant literatures.

This thesis is about the psychology of morality: it is a study of how human minds work in moral contexts. As aforementioned, in this thesis, I ask the question of whether we have an innate psychological faculty for moral judgement. To investigate this question, I will take an interdisciplinary approach: while I will refer to ethical theories in moral philosophy which provide claims on human moral psychology, I will also look at empirical studies in psychology, neuroscience, anthropology and archaeology.

I am not the first philosopher to show interest in human moral capacity, or to understand human nature by studying human moral psychology and behaviours. Western philosophy has a long tradition of examining human nature through the lens of human moral capacity. This tradition can be dated back to Plato and Aristotle, both of whom have extensively discussed human moral nature in relation to wellbeing and happiness, and how emotions or reasoning play a part in human moral nature. The most famous and influential philosopher to investigate this topic is arguably David Hume, whose moral philosophy and emphasis on the role of sentiments in moral judgements have a huge influence on the contemporary debate in moral philosophy in general. Not only western philosophers care about human morality. For example, back in ancient China during the late Zhou period (around 1000 B.C.- 221 B.C.), there are different theories on the moral nature of human beings (*xing* 性), and whether we have an innate virtuous psychological nature. Confucian philosopher Mengzi (孟子), who is referred to as the ‘Second Sage’ of Confucianism, has famously argued that humans are born with an innate moral disposition of being virtuous (*shan,善*):being morally good is human nature, just like water having the downward flowing tendency. Based on this core commitment, Mengzi proposes four cardinal virtues of human beings: benevolence (*ren 仁*), righteousness (*yi, 义*), propriety (*li,礼)*, wisdom (*zhi,智*). In contrast, another Confucian philosopher Xunzi argues that while human beings have the capacity of being morally good, human nature is evil, which is the opposite of virtue: the impulse of our human nature (*xing 性*) will lead us to selfishness and to act harmfully toward others(Csikszentmihalyi 2020).

I am also not the first person to take an interdisciplinary approach to understanding human moral psychology. Researchers have been exploring the psychological architecture of moral judgement since the birth of psychology (e.g., James 1891; Turiel 1983; Kohlberg 1958; Haidt, McCauley, and Rozin 1994). Entering 21th century, more studies have been conducted and more theories have been proposed to describe and explain human moral psychology (Malle 2021). To marry empirical results and traditional moral philosophical debates seemed like a promising way in which we can pinpoint the nature of moral cognition. However, when it comes to whether there is an innate moral faculty, the debate has not been settled by new empirical evidence yet, as the interpretations of empirical data vary.

Four empirical research areas and two theoretical frameworks, in particular, have influenced recent research in moral psychology. The first empirical influence comes from the empirical study of emotions, especially on the nature of disgust (For review, see Inbar and Pizarro 2021). Some researchers argue that empirical studies suggest that emotions are essential to moral judgements (e.g. Inbar, Pizarro, and Bloom 2009; Schnall, Benton, and Harvey 2008; Schnall et al. 2008; Rozin, Haidt, and McCauley 2008; Wheatley and Haidt 2005)[[1]](#footnote-2), and our moral judgements are fundamentally shaped by emotions (for objections, see: Landy and Goodwin 2015; Kelly 2011).

The second empirical influence comes from the study of human linguistic knowledge, especially Chomsky’s linguistic theory (Chomsky 1986; Pinker and Bloom 1990). Noam Chomsky famously argued that knowledge of language is possible because human beings have an innate domain-specific mechanism that deals with linguistic learning and performing. Drawing from Chomskyan linguistic theory, some researchers argue that the human moral system shares fundamental similarities to the human linguistic system. Just like how we acquire linguistic knowledge and how we entertain it, we apply similar but independent cognitive structure to acquire and entertain moral norms. This approach is also known as linguistic analogy, and those who employ it apply a similar argumentative strategy as Chomsky’s to argue for an innate moral faculty (Dwyer 1999; 2006; Mikhail 2007; 2011; Rawls 1971).

The third empirical influence comes from the study of the human rational learning system and decision-making behaviours (Xu and Tenenbaum 2007; Tversky and Kahneman 1983; 1974; Kahneman 2011; Crockett 2013; Cushman 2013). For example, Bayesian learning has recently been applied to understand moral norm learning (Nichols 2021; Partington, Nichols, and Kushnir 2023). More generally, empirical research on economic theory and game theory (J. Von Neumann and Morgenstern 1947; Ross 2005) has also influenced some moral theories. For example, some moral psychologists argue that moral cognition emerges for facilitating social cooperation and assessing whether others are good candidate for cooperation (Baumard, André, and Sperber 2013; Sperber and Baumard 2012; Fitouchi, André, and Baumard 2022; Curry 2016).

The fourth empirical influence comes from developmental psychology, especially studies on social learning in infants. This includes studies concerned with theory of mind in infants (Baron-Cohen, Leslie, and Frith 1985; Baillargeon, Scott, and He 2010), early moral evaluations in infants (Hamlin, Wynn, and Bloom 2007; Kiley Hamlin et al. 2013), and social learning in infants (Cesana-Arlotti, Kovács, and Téglás 2020). Studies on infants are referenced by moral psychologists for different purposes, but mainly moral psychologists treat studies on infants as evidence for an innate moral capacity (Hamlin, Wynn, and Bloom 2007).

These are four main empirical research areas that have had impacts on theorising in moral psychology. Of course, what I have listed above do not exhaust the empirical recourses for the study of moral psychology. For example, studies on psychopathology (Blair 1996; 1997; 2017) have been extensively cited when examining the relationship between emotions and moral capacity. Meanwhile, each field of empirical study does not always independently influence theorising about moral cognition. For example, Joshua Greene has drawn on both studies on emotions and theory of decision-making to develop his dual-process account of moral judgement, in which he argues that we have two modes of moral thinking: one is intuitive, effortless, inflexible, and driven by emotions, while the other is deliberate, effortful, flexible, and driven by reasoning (Joshuad. Greene et al. 2001; Joshua Greene 2014).

Apart from being inspired by empirical research, two existing theoretical pictures also influence the study of moral cognitions. The first theoretical framework is from a traditional debate in moral philosophy: moral rationalism versus moral sentimentalism, especially sentimentalism in the tradition inspired from David Hume. The Rationalism vs sentimentalism debates have been influential when moral psychologists try to provide their psychological accounts of moral judgements or try to design their own experiments. Some moral psychologists are moral philosophers themselves, and commit to some certain metaethical claims (Nichols 2004; Prinz 2007; May 2018). For example, Jesse Prinz is a proponent of moral sentimentalism, which argues that moral judgements are emotional in nature, while Joshua May argues that reasoning grounds moral judgement.

Another theoretical picture for moral psychologists is evolutionary psychology (Tooby and Cosmides 2005; Cosmides, Guzmán, and Tooby 2018; Kelly and Setman 2021; Henrich et al. 2022; Henrich, Heine, and Norenzayan 2010). Current moral psychologists tend to build up their theory based on the hypothesis and assumption that humans are social animals, and moral capacity is a result of evolution. Evolutionary psychology in general is a theoretical framework for explaining how humans behave and think. It proposes that our psychological mechanisms are products of natural selection: most human behaviours can be explained by our psychological states scaffolded by internal psychological mechanisms. Our unique psychological mechanisms are results of natural selections: in order to survive and adapt to the challenging environment, these psychological mechanisms have helped our ancestors to get around the world. Our moral capacity is probably one psychological mechanism that evolution gives to us. Normally, evolutionary psychologists treat our brain as a computation device which is sensitive to the environment, and is able to extract information and generate behaviours that are helpful for human beings to survive and reproduce (Tooby and Cosmides 2005). Some moral psychologists argue that it is essential for humans to *cooperate* for survival and reproduction, hence moral capacity evolved to facilitate better cooperation (Sperber and Baumard 2012; Curry 2016; Fitouchi, André, and Baumard 2022). This theory is also known as cooperation-based mutualistic moral theory, which centres the role of cooperation in moral cognition. They differ in terms of whether human cooperation can be cashed out in one single psychological mechanism, but arguably all of the cooperation-based theories stay neutral in sentimentalism/rationalism debates. Moreover, another distinct feature for moral evolutionary psychologists is that they look at empirical studies in anthropology (Cameron, Payne, and Knobe 2010), archaeology (Fitouchi, André, and Baumard 2022), and also cross-cultural studies (Henrich et al. 2022).

Hence, when cognitive scientists try to tackle moral cognition, they blend their different interpretations of empirical findings with theories from evolutionary psychology and moral philosophy to promote both new models for psychological accounts of moral cognition and new debates on the psychological mechanisms for moral judgements. The process of orchestrating different elements from empirical and theoretical research is complex and various, which results in different models for moral cognition. In my thesis, I will look at some of the dominant and popular accounts. These accounts include two contemporary sentimentalisms proposed by Jesse Prinz (Prinz 2007) and Shaun Nichols (Nichols 2004). Both philosophers have been inspired by the moral sentimentalism tradition, therefore they both centre emotional response in human moral cognition, and refer to studies from developmental psychology (how infants make moral evaluations, for example) and studies on psychopaths to defend their sentimentalist views. They both argue that moral judgements do not require a complex cognitive process to achieve. At the same time, both researchers take a descriptive approach to understanding moral cognition and do not use evolutionary psychology as their theoretical toolkits. However, there are significant differences in their views when it comes to the scope of moral emotions and the relationships between emotions and reasoning in moral cases. Their differences can be seen in their different interpretations of empirical data. I will discuss their views in detail in chapter 1 and chapter 2.

Two other accounts I will evaluate include a moral nativism account of moral cognition, Universal Moral Grammar proposed by John Mikhail (Mikhail 2007), and a cooperation-based moral theory called mutualistic theory of morality (Baumard, André, and Sperber 2013). This line of moral theories is heavily influenced by research in evolutionary psychology, especially how human behaviours are shaped by the motivation to adapt to the environment. While they do not commit themselves to any metaethical claims (for example, they are not particularly interested in whether emotion or reasoning is the source of moral cognitions), they contend that moral cognition is a kind of human cognition which emerges due to evolution. Hence, our moral system is sensitive to both biological and cultural environments. However, their strategies differ: Chomskyan moral nativist accounts including John Mikhail’s Universal Moral Grammar has been influenced by Noam Chomsky (Chomsky 1986) and John Rawls (Rawls 2020), both of whom have make an analogy between moral cognition and human language. Hence, the most important argument that Mikhail and other moral Chomskyans have provided is known as *linguistic analogy*, and their model is known as *Chomskyan moral nativism.* Cooperation-based moral theory rests on the assumption that moral cognition emerges due to the drive to cooperate with others. Unlike Chomskyan moral nativists drawing an analogy between the moral system and other cognitive systems, especially language, cooperation-based moral theorists conceptualise the moral system in relation to the needs to cooperate with others.

The third kind of theory I will assess is called the rational rules account of moral cognition, proposed by Shaun Nichols (Nichols 2021). Similar to Prinz’s account and Nichols’s own sentimental rule account, the rational learning account also recognises the significant role of emotion in moral cognition, and how moral cognition is activated by interacting with others and the environment. However, the rational learning account emphasises sensitivity to evidence in the environment, rather than focusing on the role of emotions. Similar to Chomskyan moral nativism and cooperation-based moral theory, the rational learning account emphasises how moral cognitions are realised in mental representation and information processing in our mind. However, while Chomskyan moral nativists and cooperation-based moral theorists tend to ground moral cognition in a unique and domain-specific mental module, the rational learning account proposes that our moral norms can be acquired via general rational learning and inferencing.

As can be seen from what I have mentioned above, while each theory agrees on some key issue (for example, all theories acknowledge that emotion is important to moral cognition, and that our moral cognition is influenced by the environment), each theory also disagrees with each other due to different interpretation of data and their theoretical grounds. At the same time, interpretations of each theory differ as well. My job in this thesis is not to settle down all the debates, as it is possible that our understanding of moral cognition is still at the early stage (although the study of human morality has lasted more than 2000 years). Instead, my job here is to provide my understanding of each theory, and to draw implications from each theory. I will argue that based on current studies and debate, I agree with Prinz and 2004 version of Nichols that emotions are important to our moral cognition. However, I do not agree that emotions ground moral judgement. At the same time, I will argue that moral capacity is not realised by any domain-specific psychological architecture. In this sense, I agree with Shaun Nichols that both Chomskyan moral nativists and cooperation-based moral theorists are wrong that we have a distinct and specific moral faculty; rather, moral norms can be acquired via general learning. However, I do not think that this general learning is the Bayesian learning process as Nichols has claimed.

1. Structure of the thesis

In the previous section, I have demonstrated the question I am concerned with in my thesis, and the accounts that I will assess and look at in the main body of my thesis. I have also proposed that there is no innate and domain-specific mechanism for moral cognition. Hence the main work of my thesis will be telling my readers why Jesse Prinz and Shaun Nichols are wrong about the role of emotions in moral judgements, why Chomskyan moral nativists and cooperation-based moral theorists are wrong to argue for an innate and domain-specific moral faculty, and why we are not Bayesian moral norms learners as suggested by Nichols.

Since I am trying to reject some influential theories, I need to further explain my standards that I use to evaluate these theories. As far as I am concerned, a good moral model should be able to explain the following four central tensions in human moral practise:

1. The good and the bad role of emotions: moral cognition is accompanied by emotion. Sometimes our moral judgements are driven by emotions (Schnall, Benton, and Harvey 2008), while sometimes our emotions also manipulate our moral judgements (Landy and Goodwin 2015).



1. The strong and the weak force of moral motivations: Moral norms are normally motivational (Rosati 2016) and we are motivated to judge and act morally, but sometimes our moral motivation is overridden by non-moral reasons (FeldmanHall, Son, and Heffner 2018).



1. The universality and the variety of moral norms across cultures: Moral norms exist universally across cultures and groups, but moral norms also demonstrate variety across cultures (Henrich 2016).



1. Agreement and Disagreement of moral norms within a moral community: while within the same society, moral norms are sometimes shared and obeyed by members of the moral community, sometimes there are always ‘black sheep’ of the community who do not adopt the core moral codes, or members in the same moral community would disagree on some fundamental moral codes in their shared moral communities.



Just by looking at these four phenomena, we can see the complexity of moral cognition. Hence, it is safe to say that we are still far away from a coherent and perfect theory of moral cognition, especially without being distracted by theorists’ own metaethical and normative ethical commitments. And again, this thesis does not aim to cover every single influential account, or provide a better account than any existing ones. What I will aim for in this thesis is to convince my readers that if we admit these four phenomena in moral cognition, we should at least come to a conclusion that there is no such thing as a single domain-specific moral faculty.

Accordingly, this thesis will be structured in the following way:

The first part of the thesis will be on the role of emotions in moral cognitions. As promised, I will look at two sentimentalists’ accounts of moral judgements. Chapter 1 will focus on the constitution model of morality proposed by Jesse Prinz. Among all the theorists I am going to evaluate in my thesis, Jesse Prinz’s account puts emotion in the spotlight the most. At the same time, Prinz argues that our emotional response is shaped by our lower-level perception. Chapter 2 will focus on the sentimental rule account proposed by Shaun Nichols. Similar to Prinz, Nichols also centres emotions in moral judgements. However, unlike Prinz, Nichols also recognises the role of reason in moral judgement. In addition, his account focuses on how affective mechanisms interact with our norm theory system, which is different from Prinz’s view which claims that emotional response in moral cases is relevant to perception of bodily change.

After evaluating both Prinz and Nichols’s theories, I will have a general discussion on the role of emotions in moral judgements in chapter 3. We can see that most contemporary moral sentimentalists are indebted to Hume’s moral philosophy as well as some of the moral philosophical debates on the nature of moral utterance from the 20th century. All moral sentimentalists are correct to realise that moral judgements are inseparable from expressions of sentiments/passions/approval or disapproval. However, I will argue that they fail to acknowledge the pervasiveness of human emotions. While emotions affect human moral judgements in a positive way, emotions also overshadow the motivational force of moral cognition as well, and sometimes we are blinded by our raw emotions, which hinder our moral cognition.

Moving on, the second part of this thesis will look at two accounts that are directly influenced by evolutionary psychology. Evolutionary psychology tries to explain human psychology and behaviours by situating the cognitive capacities in evolution: some essential human cognitive capacities are results of evolution. The two models I will examine is Chomskyan moral nativism and Mutualistic cooperation-based moral theory. These two models do not commit to either the rationalism or sentimentalism camp, rather, they focus on how humans come to having the capacity of moral cognition due to reproducing and surviving pressure.

Accordingly, in chapter 4, I will look at Chomskyan moral nativism, which argues that just like how we acquire first language, we have a language-like moral grammar that is responsible for moral cognition. This includes picking out moral norms in cultures without being explicitly told what exactly the moral codes are, and that we demonstrate moral performance at an early stage. I will argue that the linguistic analogy between language and morality doesn’t work, because unlike linguistic faculty, human moral faculty does not demonstrate the feature of domain-specificity and the existence of universal moral grammar. Moving on, chapter 5 will focus on cooperation-based mutualistic moral theory, which argues that moral cognition is grounded in concerns for cooperation, which is crucial to human species. I will argue that the connection between cooperation and morality is not direct. Firstly, there are many non-moral rules for cooperative concerns, which means that cooperation can emerge different cognitive systems. Therefore, to ground moral cognition in cooperation cannot tell us more about the unique mechanism of moral faculty. Secondly,

After rejecting sentimentalists accounts and nativist accounts of moral cognition, I will look at an alternative model: the rational learning account of moral cognition proposed by Shaun Nichols. This account argues that humans can acquire moral norms through Bayesian inferencing. Similar to Chomskyan moral nativists and cooperation-based moral theory, Nichols’s approach to understanding moral cognition does not constrain in the rationalism/sentimentalism dichotomy, and instead it investigates the acquisition of moral norms. However, they disagree on whether human moral competence suggests that we have a unique and domain-specific moral faculty. While I agree with Nichols that moral norms do not require an innate and domain-specific psychological module, I do not think that the general learning mechanism for moral norm learning is Bayesian learning. This is because Bayesian learning implies that learners are rational, which is not well supported by empirical evidence. In my final chapter, I will explore what current state-of-art Large Language Models can tell us about moral cognition. If I am right that moral cognition can be achieved by general learning mechanisms, machines should be able to engage in moral cognition. I will argue that machines do indeed engage in moral cognition in a significant way.

From the outset I wish to clarify that my aim in this endeavour is not to mount an argument against an interdisciplinary approach to understanding morality; rather it is to highlight the need for sharper philosophical analysis in the process of interrogating empirical data. Appealing to empirical studies is an essential tool for gaining a deeper understanding of the factors underlying our moral judgements, but how such studies inform one’s philosophical inquiry and the conclusions that are drawn from such research should be met with deeper philosophical scrutiny. Ultimately, I will propose that more work needs to be done in order to bridge the divide between philosophical views about moral judgement and psychological empirical data.

# Part 1: Moral Sentimentalism and the Role of Emotion in Moral Judgement

In the first part of my thesis (chapter 1-3), I will look that a dominant approach of understanding the psychology of moral judgements: emotions are essential to moral judgements. This view is also known as moral sentimentalism. I will look at two prominent accounts: the first account is known as constitution sentimentalism proposed by Jesse Prinz, and the second account is sentimental rules account proposed by Shaun Nichols. Both of them have provided their respective understandings of emotions, and refer to empirical studies to justify their sentimentalist views. Hence, while unpacking their theories on moral judgements, we can also see their understanding of emotions.

Before looking at both accounts, it might be useful to have a clarification on the kind of moral sentimentalism I concern in my thesis, because moral sentimentalism can refer to different kinds of thesis in moral philosophy. As a metaethical position—the question of what is the nature of moral properties-- moral sentimentalism has been developed and defended by philosophers. In general, there are two fundamental ways:

1. Semantic thesis: to understand the nature of moral judgements by focusing on moral statements one makes.
2. Psychological thesis: to understand the nature of moral judgements by focusing on what psychological mechanisms underlie moral judgements.

If you take the approach (a), the question concerning should be: what is the nature of moral language? One could assume that moral utterances convey our moral thoughts and our moral values. Moral philosophers such as A.J Ayer (Ayer 1936), Charles Stevenson (1944) and others scrutinise the nature of moral language: most of their work on moral judgments concerns how the syntactical and semantical features of moral statements bear speakers’ moral thoughts. For example, R.M Hare argues that moral statements are prescriptive but not descriptive (R. M. Hare 1970). According to Hare, when a person makes a moral judgement that ‘it is wrong to kill an innocent person’, the word ‘wrong’ has evaluative meaning and it makes the sentence convey ‘commendatory force’ so the moral sentence is essentially in the imperative mood. Hence, according to Hare, when a person says that ‘it is wrong to kill an innocent person’, they mean ‘Do not kill an innocent person!’.

In my thesis, I remain neutral on whether semantic investigation of moral judgements is a good approach. Hence in my thesis, and my question is a psychological one (question (b)): to understand moral judgement as a human psychological capacity and to investigate the mental structures that realise it. Let me provide an analogy for better understanding. Suppose we are interested in ‘money’. We can definitely try to investigate the nature of money by examining how lay people use the term in their daily life, how the term is used by economists, *or* what is going on in people’s mind when people are engaging in economic behaviours. It is possible that one approach might be better than others, but I think there is no competition between each of the approaches, because it is possible that different approaches provide different perspectives, and will help us understand the object we inquire ultimately.

Returning to the question of moral sentimentalism. Since what I am interested in is the psychological architecture of moral judgement, and my question will not concern moral language. So I will not look at debates on semantics of moral judgements. Instead, what I will focus on are two theses which argue that emotions are essential for moral cognitions.

## Chapter 1 Worries for Constitution Sentimentalists’ Account of Moral Judgement

In chapter 1, I will look at Jesse Prinz’s constitution account of moral judgements. in moral psychology, Sentimentalism is the view that emotions are essential for moral judgement (Nichols 2004; D’Arms 2014; Slote 2010; Prinz 2006). However, in the family of sentimentalist account of moral judgment, there are different positions on the role of emotions in them. This chapter will focus on one of the leading sentimentalist models of moral judgement, namely the constitution model proposed by Prinz. According to this view, emotions are essentially constitutive of moral judgement (Prinz 2004; 2006; 2007; 2012; 2016). Prinz has claimed: ‘If emotions are parts of moral judgements, then one cannot make a moral judgement without having an emotional state’ (Prinz 2016, 52). According to Prinz, emotion, which is a non-cognitive psychological state, is the perception of patterned bodily change (Prinz 2004, 49). Moral emotions are sentiments of approval or disapproval. And they are constitutively related to such emotions as anger and shame.

This chapter will assess and evaluate Prinz’s account of moral judgement. My intention here is to persuade my readers that Prinz’s accountis mistaken. Several tasks need to be done to achieve my goal. The first mission is to articulate the constitution account itself: what is constitution sentimentalism, what is emotion within the constitution model’s paradigm, how do constitutive sentimentalists understand the relationship between emotions and moral judgements, and what arguments has Prinz provided to justify his account.

Alongside the articulation of this account, I will also need to evaluate Prinz’s arguments: 1. Whether he has provided convincing philosophical arguments to justify his claim; 2. Whether his argument is compatible with the psychological evidence he refers to; 3. And whether constitution sentimentalism has the explanatory power that Prinz has asserted. I will argue Prinz has failed to pass this tasks, concluding that constitution model is not a good model for moral judgements.

This chapter will be constructed as follows: in chapter 1, I will explain Prinz’s account of emotions and moral emotions. This step is necessary as Prinz sees emotions as playing an essential role in moral judgements. In chapter 2, I will articulate Prinz’s constitution sentimentalism; and, alongside of this, I will evaluate his philosophical arguments and empirical evidence he has provided to justify his claim. My discussion will mainly be broken down into two parts: the first part is about Prinz’s philosophical argument. Prinz has created a thought experiment called ‘Moral Mary’, as a counterfactual argument to prove that moral concepts can only be acquired through emotions. The second part of the assessment is about examining the empirical data Prinz has referred to extensively. After the discussion of his arguments, section 3 is about whether Prinz’s account has the explanatory power that Prinz has envisaged, which answer is no. In section 4, I will conclude that the constitution model is not well supported by evidence.

* 1. Preliminary Discussion: Prinz’s View on Emotions and Moral Emotions

So far there is no consensus on the nature of emotions (Izard 2010). Thus, for moral philosophers and moral psychologists, it is essential to give an account for the nature of emotions and the nature of *moral* emotions, especially for sentimentalists such as Prinz. Because for sentimentalists, they not only centre emotions in moral motivation or moral value, but also claim that emotions at least partly constitute moral judgements (D’Arms 2014, 267), or that emotions are necessary for moral judgement (Nichols 2004).

* 1. Prinz’s View on Emotion

Traditional perceputalists such as James and Lange argue that emotion is the perception of bodily change. This account is understood to be as a non-cognitive account of emotion, as they consider emotion as the perception of bodily change. According to James, we feel sorry because we cry, angry because we strike, afraid because we tremble (James 1884, 190). Thus, it is not that we cry because we feel sorry, rather, it is the opposite: ‘our feeling of (bodily) changes as they occur is the emotion'(James 1884, 189). For example, when we see a snake, our bodies will have certain reactions such as our faster heart beating, trembling, sweating and so on. The perception of these bodily responses is the constitution of fear.

However, this account faces a series of objections and doubts, one of which is that it fails to explain the causal power of emotion or the motivational role of emotion in actions in general. For example, when we see a snake and feel fear, normally we will run away. Intuitively, the action of running away is generated by our fear of the snake. The perceptualists’ account has failed to explain this: why does emotion affect actions and decisions in a certain way? Just as Dewey has observed (Dewey 1895), if emotions are mere feelings and perceptual realisations of our bodily change, how can emotions be so important?

In debt to this perceptualism tradition and also fully aware of the problems of the original James-Lange theory, Prinz (2004), like James and Lange denied the cognitive theory of emotion, which is the view that emotions comprise evaluative judgements understood as cognitive(Lazarus 1991; Nussbaum 2001). Meanwhile, Prinz argues that there are two components of a dedicated input perceptual system: specialised transducers and mental representations. Prinz maintains that emotions are the perceptions of bodily changes, while further arguing that what we perceive is not the just the object, but the formal object, which is the mental representation of the stimulation. For example, when you see a snake, this will make you tremble and fear. The feeling of fear is composed of two elements: one is your perception of bodily changes (trembling), and the other is your perception of what he calls ‘the formal object’ of that snake, which is a mental representation of ‘danger’. Hence, the perception of snake does not constitute your emotions of fear. The mental representation of ‘danger’ will further motivate you to initiate actions (to run away).

Prinz’s account of emotions is able to avoid the motivational problem that James’s old theory faces. Moreover, this move also paves the way for Prinz to explain the motivational role emotions play in moral judgements and moral behaviours. We will discuss this in detail in section 2.

* 1. Prinz’s View on Moral Emotions and Moral Sentiments

In general, moral emotions are emotions that are either constitutive of moral judgements or causally relate to moral judgements (Prinz and Nichols 2010). For most of the sentimentalists, moral emotion is a type emotions, and distinct from other emotions as moral emotions only respond to morally significant events (Haidt 2003, 853). In contrast to this view, Prinz argues that moral emotions are derived from non-moral emotions, but there is no specific kind of emotions that are moral emotions, which exclusively serve as promoting moral behaviour and influencing moral judgement (Prinz 2012, 535). In other words, for Prinz, no emotions function uniquely as moral emotions.

Let us consider anger as an example. Anger can be triggered by different causes. In general, anger can be understood as an emotion which is evolved to bargain for better treatment (Sell et al. 2017). But for moral psychologists, anger is triggered by violations of autonomy norms (Prinz and Nichols 2010). For example, imagine these two scenarios: in the first scenario, you are angry with someone who has stolen things from a shop; in the second scenario, you are angry with your broken printer when it just stops functioning while you need to print something urgently. In the first scenario, you are making a moral judgement while expressing your anger: you are judging that it is morally wrong to steal things from a shop. In the second scenario, however, you are merely angry at the dysfunction of the printer, and it would sound weird to say that you are morally judging your printer. Comparing these two angers in moral and non-moral scenarios, we can see that there is no difference between moral anger and non-moral anger: they both elicit negative emotions towards others, and an expectation to avoid these two scenarios happening again. In this sense, they play the same functional role in both moral and non-moral scenarios. Thus, as the emotion of anger in general, it is not necessarily generated by injustice.

Thus, Prinz considers moral emotions to be pairs of emotions that respond to moral events: one is towards the action that is done by others, and the other is towards the same type of action but done by yourself. Thus, moral emotions are not defined by the types of emotions, but by the certain emotional response we have to specific actions.

It should be noted that by saying that moral judgement is emotional, Prinz does not mean that all the emotions contain moral judgements(Prinz 2007, 57). Rather, moral judgement is the emotion that ‘derives from moral sentiments’ (Prinz 2007, 64). Moral sentiment is a key concept for the architecture of constitution sentimentalism. A moral sentiment is a *disposition* that causes us to feel emotions of other-blame and self-blame towards action types, traits, and so on. In other words, a moral sentiment should at least consist of two bidirectional dispositional emotions: one is towards others and the other is towards yourself. If you only feel disgust about others doing action A while you do not feel ashamed if you did A, then you are not making a moral judgement when you feel disgusted about what others did.

Together with Prinz’s view on emotions, we now have a clearer picture of his account of moral emotions (sentiments). Moral emotions, in nature, are perceptions of the patterned bodily changes when you perceive certain types of actions. What you have perceived from those types of actions are not just bodily changes, but also a formal object of moral rightness/wrongness. This will be manifested as moral sentiments, which are the dispositional expressions of moral emotions and moral judgements.

* 1. Constitution Model Review

The constitutional model is the view that emotions are constitutive parts of moral judgement, namely to judge something morally wrong, is to (at least in part) have a negative feeling towards it:

‘The (constitutional) sentimentalist thesis asserts that, when we judge that something is wrong, one or another of these emotions will ordinarily occur, and that the judgment will be an expression of the underlying emotional disposition. (Prinz 2006)’

Therefore, in contrast to other sentimentalists’ accounts, Prinz does not think that emotions are the inputs or outputs of moral judgements, but that they *are* moral judgements. We will feel anger/disgust towards other agents if they conduct the morally wrong behaviours, while we will feel ashamed if we ourselves committed morally wrong actions. At the same time, moral emotions literally elicit moral values: feeling outrage towards an action entails the disapproval of that action. In this sense, moral judgement is an emotional state, and it is defined by the valence of that emotional state. Meanwhile, Prinz’s account does not deny the role of reasoning in moral judgements, as moral sentiments can be formed by either emotions or reasoning about the events, and reasoning can ‘uncover facts toward which we have prior sentiments’ (Prinz 2016, 66). For example, when you judge that the criminal justice system is unfair to African Americans by rationally doing research, reading articles and so on. After that you feel outrage about the government, initiating the moral emotions of anger and making the moral judgement of wrongness. In this sense, reasoning serves as one of the important elements of forming moral sentiments. However, unlike emotions, reasoning is not *directly* engaged with moral judgements. While moral sentiments can be formed by reasoning, emotions is not. ‘Reasonings can be powerful instruments in determining when we will experience our moral passions’ (Prinz 2016, 66). Thus, Prinz admits that reasonings can guide us to reach certain moral emotions but it is not a component of moral judgements.[[2]](#footnote-3)

This section will focus on the general account of constitution model and Prinz’s arguments for it. I will evaluate Prinz’s account on three main dimensions: his philosophical arguments, the empirical data he has referred to, and the explanatory power of his account respectively.

2.1 Prinz’s Philosophical Arguments for the Constitution Model

Prinz claims that a judgement must have the capacity to generate both self-blame and other-blame emotions in order to be a moral judgement. Prinz also calls this the ‘disapprobation spectrum’ (Prinz 2006). In this sense, something that merely causes you disgust cannot suffice to be morally wrong. In order for you to judge that action is morally wrong, you need to have the tendency to feel ashamed if you perform it. Thus, the strategy for Prinz is to consider moral sentiments as emotional dispositions that can be manifested in two different emotions to the same event: one is self-directed and the other is other-directed. For example, a heinous serial killer is killing an innocent person, if you are not the serial killer, you will feel angry with the wrongdoer. However, if you are that serial killer and you are still a mentally normal person, you will feel ashamed about what you have done. In this sense, emotions such as disgust, anger and shame cannot be individually counted as moral emotions, as moral emotions come in pairs and are imbedded in the moral properties of different events. In other words, Prinz also identifies moral properties as moral sentiments of approbation/disapprobation. Even though Prinz claims that the term ‘sentiment’ here refers to ‘designate an emotional disposition’ (Prinz 2007, 91), we can understand that this actually means two directions result in two different emotional outcomes. Furthermore, Prinz argues that there is little evidence suggesting that moral properties are independent of our mind and emotions (Prinz 2007, 49). In other words, moral properties (moral rightness and wrongness) are defined by our mental representation of actions and our emotional responses to them.

In line with his thought, Prinz has provided several philosophical arguments to justify his claim. In this chapter, we will have a closer look at two of them: an argument from moral development in children and the Moral Mary thought experiment, respectively.

2.1.1 Development of Moral Norms

Prinz (2006, 2007) first asks us to consider the constitution model intuitively by considering moral development in childhood. He points out that parents tend to manipulate their children’s emotions by praising, punishing and withdrawing love to teach their children moral concepts. For example, when children do something impermissible, their parents tend to punish them by withdrawing love (‘I won’t love you if you do this again!’), and children will feel sad and distressed. This process helps children to associate negative emotions with wrongdoings, further forming the moral concept of wrongness (Prinz 2006).

One might ask how we acquire the mechanism of moral norms in the first place. Prinz’s postulation is that the emotional responses we have formed are the results of moral development through emotional regulation at the early age of our lives. Bad behaviour will trigger punishment from the authority, and punishment will regulate moral behaviours in us. Gradually, morally wrong behaviours will be connected to negative sentiments, leading us to express negative emotions and promote prohibitions.

However, this argument cannot prove that this is how we develop moral capacity. In reality, there are other ways in which children learn moral values. First of all, there are empirical findings suggesting that it is possible that morality is innate. For example, Hamlin, Wynn and Bloom (2007) have shown that 6- and 10-month-old infants have the cognitive capacity to evaluate the actions of helping and hindering. In the experiment, they ask infants to watch a set of videos in which there are three cartoon agents with different shapes and colours: red round, yellow triangle and blue square respectively. In the videos, the yellow triangle consistently helps the neutral agent red round in several scenarios, while the blue square always tries to hinder the red round from climbing to the top of the hill. After the videos, infants are presented with three pieces of physical toys which are identical to the three cartoon agents that appeared on the videos, and they need to choose which is their favourite toy. As it turns out, infants prefer the helper to the neutral, and prefer the neutral to the hinderers. This study vividly shows that humans have the capacity to morally evaluate actions at an early age. No parent starts teaching their children morality when their children are preverbal infants, and no infant needs capacity to learn morality through their parents’ emotional manipulation. This suggests that children do not need to acquire moral concepts through empirical practice and learning. Instead, children innately have moral values of right or wrong without emotional regulations.

Secondly, even if children do learn some moral concepts through education, punishment is not the major strategy educators use. Rather, patient guidance and encouragement and showing empathy are more commonly used compared to mere punishment. Moreover, it is not predictable that punishment necessarily leads to the desired outcome: it is possible that children will become resentful or develop a violent inclination. Thus, punishment may not be a good way to facilitate moral value. While admittedly it is an efficient way to punish the wrongdoers, showing the moral outrage of the third-party or the victims, punishment is not a reliable way to link emotion with moral judgement. Thus, this argument fails to explain the development of association between emotions and moral judgements.

2.1.2 The Acquisition of Moral concepts

Apart from arguing that moral norms are developed through emotions, Prinz contends that moral concepts of wrongness and rightness essentially involve emotions. Prinz (2007) creates a thought experiment called ‘Moral Mary’ to argue that moral concepts cannot be acquired without emotions. This thought experiment is adapted from Frank Jackson’s famous thought experiment on the knowledge argument (Jackson 1986). In the original thought experiment, an intelligent neuroscientist Mary was brought up in a black-and-white room. She has completed her lesson on the neuroscience of colour processing in humans, and all physical knowledge about colour. However, she still lacks the knowledge of what it is like to see colours. Thus when she walks out off the black-and-white room and sees red for the first time, she will learn something new. Jackson intends to use this thought experiment to argue against physicalism, claiming that physical knowledge cannot exhaust all the knowledge in the world, and there are properties, such as redness, that are not physical properties. In the light of this thought experiment, Prinz creates its moral Mary thought experiment:

‘Imagine a woman named Mary who was never exposed to any moral education while she was growing up, but her other cognitive capacities developed normally. She is now an intelligent adult. Imagine that Mary has no intact innate moral attitudes. She doesn’t feel guilty or indignant about anything. But she decides that she wants to learn what morality is all about, so she coops herself up in a room with masterworks by Kant, Mill, and other normative ethicists. She learns their theories, and she becomes very adept at identifying the kinds of considerations that they bring to bear. For any action that she considers, Mary is able to determine (a) whether it would maximize utility and (b) whether it would lead to any practical contractions if it were pursued by all agents. Indeed, she can discern any of the facts emphasized by leading normative theories. Now here’s the crucial question. Suppose Mary discovers that doing X will in fact maximize utility. Is that sufficient for her knowing that doing X is morally right? Can she wonder whether X is morally required even though she knows that it maximizes utility?’ (Prinz 2007, 38)

Prinz intends to utilise this Moral Mary thought experiment to show that moral emotions essentially constitute moral concepts. Let us reconstruct his argument in the

If emotions are not essential components of moral judgments, then Mary is able to make moral judgments through the use of reason alone. But, Mary is not able to make moral judgments through the use of reason alone. So, emotions are essential components of moral judgments, Prinz concludes that (Prinz 2007, 39):

1. Moral concepts cannot be acquired through reasoning without emotions, thus
2. Moral concepts can only be acquired through emotions.
3. Moral concepts are essentially involved in moral judgements, thus
4. Emotions are necessary for forming moral concepts and moral judgements, and judgements based on pure reasoning cannot count as moral ones.

Prinz finds it intuitive to make an analogy between colours and moral concepts. In Prinz’s account, like colours, moral values and moral norms seem to be natural kinds out there, but in fact, they are features of our experiences. Moreover, as mentioned above, Prinz identifies morality with moral sentiments, and moral sentiments are also perceptual experiences of certain bodily changes. Thus, following Prinz’s account, it makes sense draw an analogy between morality and perceptual experience of colour .

However, there are several flaws in this thought experiment. The first is the claim that we cannot make moral judgements through the use of reason alone, which leads Prinz to his conclusion that emotions is everything being wrong. Moral reasoning can be more important than his thought, and moral reasoning can be either conscious or non-conscious (Nisbett and Wilson 1977a; 1977b; May and Kumar 2018). It is possible that this unconscious mental process is not accessible to introspection, as Kozuch and Nichols (2011) have argued that people tend to overestimate the amount of mental events they can know through introspection. For example, when solving a difficult mathematics problem, you think so hard and all of the sudden you get the answer. When people ask you how you came up with this strategy, you might reply: I don’t know. However, when you cannot explain your mental action while solving the problem, it does not mean that you figure it out by emotion. In the case of moral judgement, I think that there is a possibility that agents fail to articulate their psychological procedures but they actually employ reasoning. The reason that emotions are frequently reported might be the fact that emotion naturally feels more intense than reasoning. Thus, it is not difficult to postulate that people can make moral judgement with unconscious reasoning. If all the reasoning elements are included, it is possible that Mary can have moral concept. So to answer Prinz’s question, if Mary is that intelligent and smart, she will have moral concepts, and she will not say that doing X is right simply by referring to what Mill has said.

To sum up, I argue that Prinz does not provide a convincing argument to justify his claim. The example on moral learning with children cannot prove that emotions constitute moral norms, because the development of our moral values are not only based on emotions. Meanwhile, the moral Mary thought experiment also fails to illustrate that we can *only* acquire moral concepts though emotions. Compared to Jackson’s original thought experiment, Prinz’s Mary cannot prove to us that the moral properties are like the perceptual experience colour. Jackson’s Mary has led us to reconsider perceptual experience and phenomenology of perception, while Prinz’s Mary fails to articulate that the perceptual experience of morality is the moral sentiments.

2.2 Whether the Empirical Data Support the Constitution Model: Psychological Experiments on Psychopathy

Apart from providing conceptual analysis and thought experiments to defend his view, Prinz also extensively refers to empirical data. To support his arguments, Prinz refers to a series psychological experiments of psychopaths by James Blair and colleagues (Blair and Blair 1995; Blair 1997; Blair et al. 1997). Psychopaths in those experiments more or less show some antisocial behaviours and have a low-level deficit in emotional response to distressing cues (Blair et al. 1997). According to Prinz, psychopaths are perfect evidence for the necessity thesis, because ‘they are profoundly deficient in negative emotions, especially fear and sadness’ (Prinz 2006, 32). If scientific evidence can prove that psychopaths fail to grasp moral knowledge and make moral judgements, it will defend the constitution thesis.

In Blair’s original experiment (1995), psychopaths were asked two kinds of judgements: moral judgements and conventional judgements. In this experiment, moral judgements refer to those normative judgements that are authority-independent, more serious and related to justice, while conventional judgements refer to a cluster of judgments that are authority-dependent and less serious. The results suggested that psychopaths fail to make a distinction between moral and conventional judgements. Moreover, psychopaths tend to characterise conventional judgements as moral judgements. Based on this result, Prinz argues that psychopaths only *seem* to comprehend morality, but they really do not (Prinz 2007, 43): the result has shown that psychopaths cannot differentiate moral judgements from other kinds of judgements. This further suggests that psychopaths, who have the impairment of emotional mechanism do not have moral concepts and cannot make moral judgements. As a result, it must be the absence of moral emotions that results in their incapacity to make moral judgements. At the same time, even though psychopaths would claim that they know what moral judgements are, we have reasons to believe that they actually don’t.

However, this interpretation is wrong. First of all, let us suppose that psychopaths are completely emotionless. It is still questionable that they cannot make moral judgements, as revealed by data collected since the time of Prinz’s writing. For example, in ‘Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction’, Aharoni, Armstrong and Kiehl (2012) had their participants categories 40 events. They informed participants with varying degrees of psychopathy that exactly half of the listed acts were pre-rated by members of society to be morally wrong, and instructed them to determine which half met that criterion. As it turns out, they did not show any significant difference from control groups, and they could identify morally wrong actions. This finding indicates that psychopaths do have moral concepts and moral knowledge, they just do not care when they act. In other words, morality and moral judgement fail to be a motivational force for their actions.

Secondly, even if psychopaths are amoral, it is hard to conclude that the lack of the emotional mechanism is the reason that they fail to grasp moral concepts. Psychological experiments and psychopathy checklists do not suggest that psychopaths are emotionless. Psychopaths are able to react emotionally: for example, psychopaths as well known for their impulsive behaviors (acts based on anger). Meanwhile, psychopaths show normal reactions to emotional stimulation such as fear. In another experiment done by Blair and his colleagues (1997), psychopaths show reduced response to distress cues, while they respond to fear cues normally compared to the control group. This finding can show that psychopaths is able to emotionally respond to the *direct* emotional cues from the environment. The key difference between psychopaths and nonpsychopaths is that psychopaths are unlikely to initiate their emotion attributions to others when the distress cues are represented.

In addition, it is worth noticing that Blair’s experiments can also illuminate that there is a cognitive difference between affective emotions and empathic emotions. For example, feeling fear directly from environmental stimulation (e.g. feeling fear after seeing a snake), is different from feeling fear from empathising others (eg. feeling fear after seeing others surrounded by a lot of snakes). If the lack of the empathic emotions are the main reasons that psychopaths do not have moral concepts, then it implies that empathic engagement is more essential than affective emotions in moral judgement. However, as mentioned before, Prinz has denied the role of empathy in moral judgements. Thus, Prinz’s argument fails to show that psychopaths lack the very emotions his account requires for having moral judgements.

* 1. The Explanatory Power of Constitution Model.

I have shown that the empirical data Prinz refers to do not support the constitution model. Prinz argues that apart from the explanatory power of interpreting the empirical data, constitution sentimentalism also plays a powerful explanatory role in solving traditional philosophical debates on moral judgement (Prinz 2006). According to Prinz, it can help us understand the relationship between moral judgement and moral motivation, and the distinction between moral and conventional rules.

3.1 Whether the Constitution Model Can Explain Moral Motivation in Moral Judgement

In the literature on moral motivations, there are debates on the nature of motivation, relationship between moral actions and moral motivations, including the question of what motivates us to act morally (Schroeder, Roskies, and Nichols 2010). According to Prinz (Prinz 2006), emotions are the intrinsic motivation for us to act morally. We act morally good because we have a positive emotional disposition towards it, and we condemn wrongdoings in virtue of our negative emotional response to the morally wrong behaviours. In other words, it is the emotion that internally motivates us to think an action is wrong, promoting further avoidance and intervention.

However, the constitution model cannot explain the relationship between moral behaviours and moral motivations. Firstly, this model cannot explain why mentally healthy people will conduct morally wrong action if moral sentiments can be the ultimate motivations for morally good actions. It is not rare that people will conduct morally wrong behaviour even though they know it is wrong to do so. For example, there is a crazy Liverpool fan who disguised himself as a cleaner, sneaking into the stadium so that he can watch an important game. He was fully aware that lying is morally wrong. Before he decided to disguise himself, there are two options for him: the first is to obey his moral principle and suppress his passion for the club; the second is to act impulsively. They are conflicted options but both options are out of his care for his happiness: the former is good for his reputation among the society, the latter is good for his enjoyment of an important football match. If emotions entail moral concepts and moral judgements, and further motivate us to act in a certain morally good way, why did he still choose to cheat, fulfilling his other emotional desire? Why will people choose to do something impulsively disregarding the moral concept they have already grasped? Why does the emotion towards moral events become inert when other emotions about non-moral events appear? It seems to me that it is because moral actions are motivated by various reasons, and emotional reasons play a less significant role in motivating moral actions than they do in promoting other kinds of actions. The emotions towards moral events are common, however, they are not as essential as Prinz has envisaged. Rather, emotions play a much more important role in other non-moral motivations.

Moreover, constitution sentimentalism cannot explain the moral motivations in altruistic actions. Let us think about Green et al.’s (2001; 2004) example: imagine a scenario in which enemy troops invaded a small village. People are hiding in a small tent, while there is a baby crying. The mother of the baby is facing a dilemma: whether she needs to smother her baby. If she refuses to do so, everyone will die. However, it will cause herself suffering if she does so. Let us further suppose that, the mother decided to smother her baby to save others’ lives because she thinks this is right to do so. She feels agony when she smothers her baby, and it is counterintuitive to think that she will have any positive emotion even though she thinks it is morally right to save people in the tent. In this example, the moral concept in the mother is not constituted by her emotional responses. If you are one of the villagers in the tent, you might think that it is morally better for the mother to smother the baby. However, you do not necessarily smother your baby if you are that mother. Thus, altruistic actions are not necessarily motivated by emotions, and even if that is the case, this emotional response is not bidirectional, and motivate the same actions just like Prinz has suggested.

3.2 Whether the Constitution Model Can Explain the Moral/Conventional Distinction

Now, let us consider Prinz’s claim that sentimentalism can solve the problem of the moral/conventional distinction. Similar to moral judgement, conventional judgement is also about judging an action permissible or impermissible, and it tends to co-occur with emotion. Thus, it draws heated debates on how can we distinguish between moral and conventional judgement. The debates have been influenced by Turiel (1983) and his followers, who argue that moral rules are taken to be authority-independent, objective, about justice and more serious. Contemporary philosophers tend to hold the view that morality is universal and impersonal (Goodwin and Darley 2008). Prinz proposes that constitution sentimentalism can explain how we come to draw the moral and conventional distinction: by appealing to different emotional responses. Prinz argues that we are less emotional about conventional rules, and that emotions might be grounded in moral rules (Prinz 2006). For example, speaking without raising your hand might violate conventional rules and makes you embarrassed, but this embarrassment is bound up with the reactions of others. If others do not mind your behaviour, this emotion will disappear. On the contrary, you will feel bad if you hit someone no matter if it is prohibited by the authority. The feeling of guilt about hitting others will not fade away even if you try your best to suppress this feeling.

However, I argue that this is counterintuitive, and the difference between moral/conventional judgements cannot be distinguished by emotions. First of all, Kelly and his colleagues (1993b) have found that there is no significantly different pattern for what people think about moral judgements and conventional judgements. For example, when participants are asked whether whipping is ok 400 years ago and now, they tend to answer that it is not morally wrong to whip 400 years ago, but it is very bad and morally wrong to whip others nowadays. This finding suggests that we view our moral judgements are authority-dependent. If sentimentalism is right, and we feel bad about whipping, why do we still judge that whipping nowadays is morally worse than 400 years ago? As far as I am concerned, the answer must be that more concern and psychological processes have been involved in making a moral judgement beyond just emotional responses. When we are making a moral judgement, we do not act just based on the emotional response to an event. Rather, before making a judgement, some other moral considerations, such as moral reasoning and reflection on the specific circumstances are involved.

Additionally, and also quite opposed to Prinz’s claims, other researchers have found that we tend to react quickly and intensively to the violation to conventional rules. Haidt and his colleagues (Haidt, Koller, and Dias 1993a) examined Americans and Brazilians’ attitudes towards actions such as eating your dead pet dog, having sex with a dead chicken, and washing your toilet with national flags. As it turns out, people will feel outrage by violating the local conventions. Thus, it is not counterintuitive to think that people tend to react emotionally to the conventional violations as well, and the intensity and sincerity of the emotional response cannot distinguish moral judgements from conventional judgements.

* 1. Conclusion

The constitution model of sentimentalism is the view that emotions are the basis of moral judgements, in the sense that emotions provide a sentimental foundation for moral judgements. Moral judgements are emotions in the sense that moral judgements are literally the manifestation of moral sentiments. However, as can be shown above, none of the arguments Prinz has provided can convince his readers to believe that emotions are essentially constituted by moral judgements. Prinz’s philosophical argument about moral development tries to illustrate that children learn moral properties through emotions, and further intends to prove that moral properties are naturally connected to emotions. However, I argue that there remains a dispute on whether we learn moral properties through emotional manipulations. Firstly, it is possible that the cognitive capacity of understanding moral properties is innate; secondly, there are more other approach in which we learn morality. Meanwhile, Prinz creates a moral Mary thought experiment, arguing that moral concepts can only be acquired through emotions. However, emotionless Mary cannot help Prinz to prove that moral concepts cannot be acquired through other psychological processes, because.

Apart from the philosophical arguments, Prinz has referred to psychological experiments to prove that moral judgements are essentially constituted by emotions. Psychopaths, according to Prinz, can vividly show that people who have impairment with emotional response fail to make moral judgements, which implies that emotions are essentially involved in moral judgements. However, I argue that he has misinterpreted Blair’s experiments with psychopaths, and psychopaths are not deficit in emotional response, thus experiments with psychopaths cannot prove that emotions constitute moral judgements. Moreover, even if psychopaths have emotional deficiency, there is evidence to show that they can grasp moral concepts. Those empirical studies can only show a relevant relation between emotions and moral judgement, rather than an essential one.

## Chapter 2 Mind the Gap: Why Empirical Studies Do Not Support the Sentimental Rules Account

It is increasingly common within the literature on normative theories of moral judgement that the findings of empirical research are used as a means of justifying one’s philosophical claims. Shaun Nichols is a leading figure within the debate concerning normative theories of moral judgment (Nichols and Knobe 2007) and is a key proponent of appealing to empirical data (Nichols 2005) to bolster the philosophical claims underpinning his position. At the heart of Nichols’s project is a commitment to the idea of moral sentimentalism, which essentially claims that emotions play a fundamental role in moral judgement (Nichols 2005; Prinz and Nichols 2010). Precisely, according to Nichols, our moral judgements are caused by our possessing a normative theory of prohibiting harms (sometimes disgust), together with our experiencing an affective response to distress cues of others’ suffering (Nichols 2004, 62). To support this claim Nichols surveys a wealth of empirical data, in particular focussing on experiments with diagnosed psychopaths and autistic children[[3]](#footnote-4) (Blair 1995; 1997; Baron-Cohen, Leslie, and Frith 1985). This chapter will examine Nichols’s position in light of his appeal to this empirical data in order to assess the philosophical worth of the sentimentalist thesis he upholds. I argue that the experiments Nichols cites ultimately failto support his view on moral judgement, that is, unless he abandons his commitment to sentimentalism.

My chapter will be structured according to three sections. The first section will provide an outline of Nichols’s position and his evaluation of the empirical data. In particular, I will unpack the experiments Nichols evaluates relating to psychopathy and child autism, which he argues illustrates the essential role emotions play in moral judgement. I argue, however, that such experiments do not directly imply that emotions are essential to moral judgement, nor that an irrational affective mechanism is essential to moral judgement. Having established these initial concerns with the core premises of Nichols’s project, the second section will provide a sustained critique of Nichols’s Sentimental Rules account by showing that it fails to pinpoint the nature of moral judgement and also fails to adequately explain judgements of moral goodnessjudgement. In the case of the former, I suggest that Sentimental Rules are not specific to instances of moral concern as they are frequently triggered in demonstrably non-moral contexts, such as sports events, for example. In the latter case, the sentimental rules account cannot explain how we judge altruistic actions. In the last section, I will attempt to assess and evaluate the methodology that Nichols uses to justify his philosophical and meta-ethical commitments; namely, his appeal to empirical data. I agree with the principle that utilising empirical data is necessary to further our understanding of how agents make moral judgements, and on that basis Nichols’s avenue of inquiry is sound. However, my contention is that the evidence he provides cannot support his claim, because those experiments cannot directly point out that emotional responses can exclusively initiate moral evaluations. The discrepancy between what the empirical findings reveal and the conclusions that Nichols draws bears significant implications for how we should interpret and utilise empirical data in research relating to moral judgement.

1. Why the Sentimental Rules Account Is Wrong about the Empirical Evidence
   1. Preliminary Work: What is the Sentimental Rules Account and Why Focus on It

The Sentimental Rules account (SR account hereafter) is a philosophical claim about what psychological capacities are the keys to moral judgement. It is a version of sentimentalism, the general position which contends that emotion is an? essential element of moral judgement. The SR account claims that there are two independent psychological mechanisms required to form moral judgement: one is a mental representation of normative theory on harm-prohibiting actions; another is a negative affective response to others’ suffering (Nichols 2004, 29). As Nichols has put it: ‘Core moral judgement thus implicates what I will call ‘Sentimental Rules’, Rules prohibiting actions that are independently likely to elicit strong negative affect’ (Nichols 2004, 18). Meanwhile, he argues that ‘both the affective mechanism and the normative theory are both developmentally necessary for achieving the nonconventional normative theory.’ (Nichols 2004, 26).

In other words, moral judgement is a manifestation of the conspiration of two distinct psychological processes: a normative theory about harm/disgust prohibition and affective response. They interact with each other but are dissociable. The normative theory is a psychological mechanism bearing the belief that harm should be prohibited—a theory which is socially and evolutionally acquired. The initiation of the normative theory about prohibiting harms on others only requires minimal mindreading capacities. This moral normative mechanism can help us distinguish moral contexts from non-moral contexts. Together with the trigger of the affective response, which is identical to emotional response, to others’ suffering, we can be able to make moral judgements. In addition, Nichols argues that because nonconventional judgements appear at the early stage of humans, these two psychological mechanisms are not advanced, and do not require very sophisticated mindreading capacities. Meanwhile, as moral judgements tend to be fast and intuitive, the mechanisms that underlie them are not any type of rationality, according to Nichols. Nichols does not say a lot about the emotional response mechanism and how it relates to normative theory of harm prohibiting. But he argues that if moral judgements imply objective moral facts, then objective purports need to be imbedded in affective response and norms about harm prohibiting (Nichols 2004, 179).

The Sentimental Rules account has recently become prominent as sentimentalism has started to dominate psychological research in moral judgement. Unlike traditional moral philosophers who examine the nature of morality by conceptual analysis, Shaun Nichols has been trying to defend his account from an interdisciplinary approach. He has expanded his argumentations and justifications through spanning the empirical data across cognitive psychology, neuroscience, developmental psychology, etc. In other words, his arguments are heavily experiments-backed, and he utilises these data not only to support his empirical claim, but also argues against other meta-ethical theses.

What Nichols has covered in his argumentations are indeed impressive. However, the more we look at the empirical data and psychological experiments, the more can we notice a clear gap between the evidence and Nichols’s conclusion. Even though Nichols says his sentimental rules proposal can best explain existing scientific data about human morality, the evidences are actually very mixed, with no sign of settling down the philosophical debate on moral judgements.

* 1. Empirical Data on Psychopathy and Autistic Children Evaluated by Nichols

Experiments with psychopaths and autistic children are mentioned frequently throughout Nichols’s work. They serve as counterexamples for Nichols’s opponents: moral rationalism, traditional sentimentalism and neo-sentimentalism, indirectly supporting Nichols’s own SR account.

Before getting into the empirical evidence, let us get a sense of the alternative philosophical claims about moral judgements. In general, moral rationalism is the view that moral judgement is grounded in reasoning (Smith 1994). Traditional sentimentalism, which is largely in debt to David Hume (Hume 1751), claims that emotions are essential to moral judgement since making a moral judgement is intrinsically linked to being motivated to judge and act in a certain way, and emotions are the best candidate for exhibiting this motivational force. Neosentimentalism is the view that to think that an action is wrong is to think that it is appropriate to feel certain emotion in response to this action (D’Arms and Jacobson 2000; D’Arms 2014). For example, to judge that stealing is wrong is to think that it is appropriate to feel angry in response to someone stealing things. Nichols argues that empirical data do not support any of these three philosophical claims. Instead, the recent scientific findings are fit better with the SR account. The empirical evidence he presents are mainly experiments with psychopaths and autistic children.

Psychopaths have long been considered as amoralists who lack the capacity to make moral judgements, as they are not motivated to act morally, and they suffer from different cognitive impairments. As Kent Kiehl puts it, psychopaths are ‘callous, shallow, and superficial, and they lack insight and empathy for the effect their poor behaviour has on others; behaviourally, psychopaths are impulsive, nomadic and have weak control’ (Kiehl 2006, 109). As consequence, while psychopaths can be a great case study for moral judgement, it is hard to determine which deficits in psychopaths are the main reason that they cannot make moral judgement (if they cannot). Thus, other empirical data are needed as valuable reference.

At the same time, the psychological studies on autism can help with that. Autism is a mental disorder which shows impairment in verbal and nonverbal social communication. People who suffer from autism tend to have difficulty communicating with others (Baron-Cohen, Leslie, and Frith 1985; Baron-Cohen 2019). However, regardless of their affective deficits, they are able to make moral judgements (Blair 1996). Thus, by comparing these two groups of studies on different subjects together, we might be able to pinpoint the capacities that are essential to moral judgement. And that is also how Nichols construct his arguments against his opponents.

Nichols’s explanation of empirical data on psychopaths is mainly based on the studies conducted by James Blair and his colleagues (Blair 1995; 1996; 1997). In different studies, they have studied the moral judgement of psychopaths (Blair 1995), children with psychopathic traits (Blair 1997), and autistic children (Blair 1996). According to these series of experiments, they found out that psychopaths tend not to distinguish moral judgements from conventional judgements. However, autistic children are able to tell the difference between moral and conventional violations. I shall unpack those experiments separately.

In the experiment on whether psychopaths can make moral judgement, based on psychopathic checklist score (R. D. Hare 1991), Blair (Blair 1995) recruited 10 psychopaths and 10 non-psychopaths as a control group in a prison in London. The psychopathic checklist score is a reliable assessment about psychopathy degree, and has been widely applied in clinical psychology and criminal assessment. The psychopathy checklist consists of 2 factors with 4 facets: Factor 1 contains the facet of interpersonal traits and the facet of affective traits, while factor 2 contains the facet of lifestyle traits and the facet of antisocial traits.

A screenshot of a cell phone

Description automatically generated

Figure Psychopathy Checklist (R.D. Hare 1991)

Blair asked participants to consider 8 school scenarios in which rules had been violated. Some of the rules were moral and some were conventional, but the participants were not alerted to this fact. They were simply asked to rate the wrongness and seriousness of those violations, and to justify their answers. Meanwhile, they were also asked whether the described behavior would have been wrong if an authority had allowed it. Blair predicts that psychopaths would not be able to identify moral judgements because psychopaths think moral judgements are convention-like, which is dependent of the authorities. And Blair also predicts that psychopaths would ignore victims' welfare. The results were striking: psychopaths did not treat moral and conventional wrongs significantly different. However, unlike a control group of non-psychopathic criminals, they tended to ignore victim’s welfare when justifying their answers about moral wrongs and psychopaths tended to treat both moral and conventional wrongs as if they were authority-independent. Blair speculates that this is due to the fact that psychopaths try to hide their inability to make the moral/ conventional distinction. Thus, Blair argues that psychopaths actually cannot make moral judgements.

Another important experiment on psychopaths is about their response to different stimuli (Blair et al. 1997). In this experiment, 18 psychopathic individuals and 18 incarnated controlled individuals participated in the experiment. They were shown three types of stimuli: threatening, distress cues of others and neutral stimuli. There are 28 slides: 10 are for practicing, 5 depicting a distress cue (for example a crying face), 5 expressing threatening (a pointed gun), and 8 on neural stimuli (a book). The result shows that psychopaths cannot properly respond to others’ distress cues while they can respond to other cues such as threatening and neutral stimuli. This has shown psychopaths have impairments in cognitive mechanisms that are required for responding to others’ emotions, such as empathy or mindreading.

The third experiment I will unpack is the classic false belief task. The false belief task is a classic method in the study of cognitive development. It has been treated as a task to test if an agent has capacity to represent others’ mental states, which is also known as theory of mind. In the original task (Baron-Cohen, Leslie, and Frith 1985), participants are presented a scenario in which there are two protagonists, Sally and Anne. Sally puts her marble into her basket and leaves the room. After Sally leaves, Anne takes out the marble and puts it into another box. After watching the scenario, participants are asked when Sally returns to the room, where will she find her marble. Baron-Cohen and collaborators found out that 16 of the 20 autistic children (80%) have failed to reply that Sally should look for the marble in the basket. This result has been replicated several times. The result has led scholars to conclude that autistic children lack the capacity to mentalise others, also known as theory of mind.

Blair further investigates morality in autistic children (Blair 1996). In this experiment, 20 children with autism participated in the experiment, and they were divided into two groups: one group (10 persons) has passed the false belief task while another cannot pass the false belief task (10 persons). At the same time, 10 normally developing children and 10 moderate learning difficulty children were chosen as control groups. The result has shown that both groups of autistic children can make distinctions between moral judgement and conventional judgement, and they perform the same as control groups.

Based on the results of the four experiments mentioned above, Nichols constructed his arguments against his opponents. Firstly, he adopts the initial conclusions from the original experiments: 1. Psychopaths lack the capacity for making moral judgements. 2. Psychopaths cannot affectively respond to distress cues. 3. Psychopaths are able to exhibit some emotions such as threatening and fear. 4. Autistic children lack the capacity of representing others’ mental states, thus they lack a general rationality, or at least completed rationality 5. Autistic children can be able to make moral judgements.

Now it’s time to test each philosophical claim. Nichols firstly argues against moral rationalism by emphasising conclusion 1, 4 and 5. According to conclusion 1 and 2, it seems that psychopaths cannot make moral judgement because they cannot empathise with others through understanding others’ mind. However, experiments on autistic children has shown that they also suffer from the deficit in theory of mind, but they can be able to make moral judgements. According to Nichols (Nichols 2004, 79), one way of testing whether an agent has general rationality is to see if she can pass false belief task as mentioned above. Since autistic children cannot pass false belief task (conclusion 4), this suggests that they have deficits in reasoning or in full-blown theory of mind. However, since they are able to distinguish moral judgement from conventional judgement (conclusion 5), there must be some psychological deficit other than reasoning that causes the failure for psychopaths to make moral judgements.

Secondly Nichols argues against other sentimentalists by emphasising conclusion 5. Nichols has pointed out that the both traditional sentimentalists and neo-sentimentalists have to answer a problem: moral judgements also appears in non-emotional contexts. In addition, neosentimentalists have to answer another question: how can some agents who lack appropriate emotions also make moral judgements? Neosentimentalists’ account requires agents to have appropriate emotions to respond to certain actions. Nichols argues that this requires agents to have certain emotions such as guilt and shame. However, young children and autistic children are not able to experience such emotions whereas they are able to make moral judgement just like normal people.

Nichols argues that the SR account is compatible with all 5 conclusions. He argues that the deficit of an affective mechanism related to responding to distress cues in psychopaths is the main reason why they can’t make moral judgements. According to Nichols, the affective response to others’ distress does not require reasoning or theory of mind, and it is not equal to empathy[[4]](#footnote-5). The affective response to others’ distress only requires a minimal level of mindreading which comes fast and intuitively (Nichols 2004, 62). Although autistic children suffer from deficit in general rationality, they still have the minimal cognitive capacity of mindreading, which is enough to enable them to affectively respond to others’ distress cues. As far as I am concerned, SR account is indeed appealing among other philosophical views. However, it is still not an adequate account to reveal the key features? of moral judgement.

* 1. The Empirical Data on Psychopathy and Autism Does Not Support SR Account

The first problem for Nichols is that the SR account is not empirically backed as Nichols has claimed (Royzman, Goodwin, and Leeman 2011). First of all, experiments with psychopaths do not directly suggest that psychopaths cannot make moral judgements. The questions asked to participants in the study were about an unfamilier subject matter or setting, which may be what explains why they were unable to differentiate moral from conventional judgments in the study(Aharoni, Sinnott-Armstrong, and Kiehl 2014). Recently, there is psychological evidence suggesting that psychopaths *do have* moral concepts and are able to make moral judgements; they just don’t care about these judgments when they behave (Cima, Tonnaer, and Hauser 2010; Aharoni, Sinnott-Armstrong, and Kiehl 2012).

For example, Aharoni and colleagues (2012) examine whether psychopaths can be able to discern moral wrongs regardless of their anti-social behaviours. Additionally, they also examine if any facet of psychopathy contributes to failing to recognise morally wrongness. There are 109 incarcerated prisoners with various degrees of psychopathy as measured by the Psychopathy Checklist—Revised score[[5]](#footnote-6) participating in the experiment. Participants are asked to assess 16 scenarios using three questions each: 1. Whether it is morally wrong? 2. How bad it is? 3. If there is no law against this act, is this still permissible?

There are 3 interesting findings in their studies. Firstly, they found out that the general PCL-R score is not the significant factor that affects the result, which means that the level of psychopathy is not associated with agents’ ability to make moral judgement (Aharoni, Sinnott-Armstrong, and Kiehl 2012, 490). As can be seen from the figure 2, participants’ score on the psychopathy checklist cannot distinguish participants’ moral accuracy. Similarly, from figure 1.3, participants who have a high score (above 25) do not perform significantly worse than students who have a normal PCL-R score.

A screenshot of a cell phone

Description automatically generated

Figure (Aharoni, Sinnott-Armstrong, and Kiehl 2012, 491)

A picture containing parked

Description automatically generated

Figure (Aharoni, Sinnott-Armstrong, and Kiehl 2012, 491)

Secondly, they found out that there is no strong evidence suggesting that any of the factors of psychopathy traits is associated with moral accuracy. As mentioned in the previous chapter, there are four different factors of psychopathy traits: interpersonal traits, affective traits, lifestyle traits and antisocial traits. They found out that none of the traits significantly affect the moral classification accuracy.

Interestingly, they have found that IQ, which is a measure of non-emotional cognitive ability, is strongly related to their performance (Aharoni, Sinnott-Armstrong, and Kiehl 2012, 491). According to the result, IQ can explain 54% of the variance in task performance. People with high IQ score can perform better in rating the moral accuracy. This experiment has shown that there is no strong evidence that psychopaths are worse at distinguishing moral from conventional transgressions, which suggest that psychopaths have the capacities to form normal moral norms. They just don’t care about moral value, or whatever concern they do have for moral value is not strong enough to motivate them to perform moral actions.

Thus, as a result, even though psychopaths cannot respond to others’ distress normally, and their affective mechanism remains idle, they can still make moral judgements. Their poor performance in moral behaviour might be the result of their deficits in general cognitive capacity, rather than the impairment of moral judgement. I think this empirical finding is much more convincing compared to Blair’s (1995). Firstly, the sample size is significantly larger compared to Blair’s, which can be more accurate to detect the correlation between psychopathy and moral accuracy classification. Secondly, the questionnaire covers much more diverse moral scenarios, rather than only focusing on school scenarios. And last but not least, Aharoni and colleagues have looked at the data in more detail. For example, they have examined whether any facet of psychopathy will affect the overall performance of participants. This can directly help us understand whether affective response, which is one of the facets of psychopathy, fundamentally influences moral judgement.

Now let us consider autistic children. I argue that it is wrong for Nichols to conclude that autistic children have deficits in general rational capacity simply based on their performance in false belief tasks. Firstly, As Bloom and German point out (Bloom and German 2000), passing the false belief task requires more cognitive capacities other than mentalising others’ minds. Meanwhile, to mentalise others’ minds does not entail the ability to reason about others’ false beliefs. To pass the false belief task, especially the classic false belief task, agents not only need to empathise others, but also to follow the actions of two protagonists in the narrative. This requires certain levels of linguistic capacities and attention. It is possible that the lack of proper linguistic and attentional resources are the main causes of autistic children’s bad performances in the false belief task. I argue that even if the majority of young autistic children cannot pass the false belief task, they still have a certain level of general reasoning capacity, including a certain degree of mindreading. Thus, studies on autism cannot justify Nichols’s view that the lack of rationality does not affect moral judgement abilities, since there is no direct evidence suggesting that autistic children lack the ability to mentalise others. Meanwhile, it is not clear what Nichols means by a minimal capacity of mindreading and why it is categorised as an emotional response. Nichols has mentioned that the minimal capacity of mindreading does not require agents to represent others’ mental states, but it does requires agents to have a certain level of cognitive capacity to respond in relation to others’ behaviours (Nichols 2004, 54). But then again, he should provide a clearer account for the affective mechanism.

In this section, I have illustrated one of the dominant sentimental accounts of moral judgement, namely Sentimental Rules account proposed by Nichols. Nichols intends to defend his account by referring to psychological experiments on psychopathy and autism. Nichols argues that these two topics can show that affective mechanism in addition to the norms of harm prohibiting are the two main mechanisms for making moral judgements: psychopaths who lack empathy and affective response cannot make moral judgements, whereas autistic children, who have impairments in general rationality but have normal function in affective response, can make moral judgements. Those results can eliminate the possibility that rationality is essential to moral judgment. As a result, Nichols concludes that sentiments are essential to moral judgement. However, in this section, I argue that empirical data from psychopathology and autism are misinterpreted by Nichols, and neither of them directly support Nichols’s claim. I have shown that the recent experiments on psychopathy and moral/conventional classification accuracy have shown that psychopaths are not significantly worse at making moral and conventional judgement compared to normal people. Meanwhile, it is not true that autistic children have impairment in basic rationality even though they cannot pass the false belief task. In the following section, I will further provide more evidence in favour of my criticism against Nichols.

1. Sentimental Rules are Not Enough: Other Evidence

As mentioned above, Nichols argues that the capacity to make moral judgement depends on two mechanisms: one is knowledge of a normative theory prohibiting harmful actions, and second, a negative affective response to distress and others’ suffering. In the previous section, I have argued that the empirical evidence that Nichols has presented does not directly support his sentimental rules account. Some of the conclusions he drew from the experiments are too coarse and inaccurate. In this section, I will further argue that there is empirical evidence and daily life experience which can further prove that sentimental rules are not an accurate descriptive account of moral judgements. There are three reasons for my conclusion. The first is that the notion of affective response is mixed. Nichols has implied that affective/emotional response is identical to fast and intuitive response. However, fast and intuitive response can be initiated without the involvement of emotions. The second is that sentimental rules occur much more often than when we make moral judgements. And the last is that the sentimental rules account cannot explain judgements of moral goodness. As a result, sentimental rules are neither sufficient nor necessary for moral judgement.

* 1. There Are More Fast and Unconscious Responses than Affective/Emotional Response

The idea of affective resonance is crucial in Nichols’s account. It composes one of the two psychological mechanisms that are required for making moral judgements. However, it is not clear what Nichols means by affective mechanism. Nichols takes it for granted that affective response is emotional response. In order to facilitate moral judgement and further influence cultural norms, affective resonance is necessary, according to Nichols (2004, 128). Although Nichols has described it as a psychological process which happens fast and involuntarily, this is not enough to distinguish it from rational processes and other non-emotional processes, which might contribute to forming moral judgements. As Joshua May points out, we can reason without effort and consciousness (May 2018). For example, when we calculate 1+1=2, it is fast and intuitive, but it would sound odd to characterise this intuitive response as emotional response.

Admittedly, emotions have strong motivational force, but it is not the case that great motivational force only comes from emotions. Reasons and social norms can have a great motivational power without emotions as well. For example, suppose that I want to write a good philosophy paper, and I am motivated to work hard everyday. My motivation is not necessarily from emotions such as fear or excitement. Rather, it is possible that I think it is my duty to write philosophy papers. This doesn’t sound as enthusiastic as emotional motivation, but the internalisation of social norms can have great motivational powers to affect our behaviour. This can also apply to moral judgements. Just as internalised social norms can operate quickly and be highly motivating, perhaps moral norms can be as well. If Nichols characterises this mechanism as emotional response mainly because it is fast and effortless, then he confuses *emotions* with *intuitions**,* and he has to provide a more solid argument on why moral norms can only be internalised and manifested in an emotional form.

* 1. There Are More Sentimental Rules than Moral Judgements

In the last subsection, I have argued that there are more unconscious, intuitive and fast responses than emotional/affective response that can be invoked for making sense of moral judgements. Some might argue that it is just an unjustified to claim that they can do this and that just as I have claimed of Nichols’s view, there is no direct evidence in favour of it. In this section and the next section I will grant that Nichols is right about the necessity of affective response for moral judgments and that affective response in morality is emotional response, and I will argue that the sentimental rules account still does not work.

Firstly, sentimental rules occur much more often than in moral judgements: there are more scenarios in which our normative knowledge of harm-prohibiting actions and affective response to others’ distress are triggered. For example, both norms about harm prohibition and affective mechanisms are frequently activated in the context of competitive sports events, especially those that contain intense physical contacts. As a long-time Liverpool supporter, I have been through numerous experiences in which my harm prohibition norm and affective response to distress were triggered, while I was not making any moral judgement. For example, in the 2018 Champions League final, Liverpool’s key player Salah was fouled by Real Madrid defender Ramos which ended Salah’s final. Also, in the same game, Liverpool goalkeeper Karius suffered a concussion after taking a blow to the head from Ramos. No yellow card was given by the referee. As a result, Liverpool lost to Real Madrid. There is no doubt that I finished watching that game with full agony and anger: I was angry about Ramos’s behaviour (my harm-prohibiting mechanism is activated), I felt very sad about Liverpool losing the match, and I also was concerned about Salah and Karius’s injuries (I have affective response to others’ distress). However, it would make me sound like an irrational fan if I insist that Ramos was doing something morally wrong. In sports, rules allow for physical contact and certain levels of harm. For example, tackles in football and rugby are common, and some players have suffered badly because of rule-sanctioned physical contact: they might not only feel the physical pain but also psychological trauma. Harms in sports might even cast shadows over player’s promising futures. Normal people will feel for them, but will not judge that morally wrong actions have been done to them in the context of sports competitions. How we react to tackle-related harm in sports events suggests that there are scenarios in which our sentimental rules are triggered while our moral judgement remains idle.

In addition, how we evaluate harms in sports events has another implication of moral judgement: it is possible that other psychological mechanisms are required in making a judgement a moral one. At least a cognitive mechanism that makes us be able to detect different harms in different contexts. This must not be part of the emotional response, but it is essential in the sense that it can distinguish moral contexts from non-moral contexts. This very capacity of distinguishing moral and non-moral contexts is hard to classify as either of the two sentimental rules that Nichols has described.

* 1. There Are More Core Moral Judgements than Sentimental Rules

Moreover, even if the sentimental rules account can explain harm-related moral judgements, it cannot account for how we judge actions that are morally good. Researchers have found that the attributions of praise and blame differ in a significant way. For example, Anderson et al. (2020) argue that while blame is for punishment, praise is for relationship building. Nichols may have provided an appealing account on how we judge actions to be morally wrong, and his analysis on altruistic behaviours seem perfectly explain how the affective response and care mechanism work together to trigger moral judgement, further motivate us to perform altruistic behaviours (Nichols 2004, 33–61). However, it does not explain how we judge actions that are morally good and that are morally permissible. I will illustrate my point through two different approaches. I will firstly present the empirical evidence by Joshua Knobe, and then I will further defend my argument by appealing to daily life experience.

Affective response to others’ suffering might trigger our altruistic behaviours, but it does not necessarily lead us to judge actions to be morally good. People show asymmetrical patterns of judgement of moral praise and moral blame. People are more motivated to identify the cause of immoral acts than moral acts (Roese and Olson 1997). As Smith has put it, there are at least three different types of moral judgement: one is to judge that an action is impermissible, a second is to judge that an action is permissible, and the last is to judge that an action is morally noble (Smith 2018). In our daily moral life, to judge an action is morally good is equally important as to judge an action is morally bad. As Joshua Knobe has pointed out (Knobe 2003b; 2003a), there is an asymmetry between blaming (judging an action is morally bad) and praising (judging an action is morally praise-worthy). In one of Knobe’s experiments, participants were randomly assigned to either the ‘harm condition’ or the ‘help condition’. Participants in the harm group are presented with the following vignette(Knobe 2003b, 191):

‘The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’

The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'

They started the new program. Sure enough, the environment was harmed.’

Participants in harm group were then asked how much blame does the chairman deserve for his statement, and whether they thought the chairman has the intention to harm the environment. In comparison, for participants in the help group, they read almost the same story, but the word ‘harm’ is changed to ‘help’ in their story. After reading the help story, participants in the harm group were asked how much praise does the chairman deserve for his decision, and whether the chairman has intentionally helped the environment.

The result has shown that two radical different patterns were elicited by two conditions: in the harm condition, most participants believe that the chairman deserves blame while in the help condition, most participant don’t think that the chairman deserves praise. This experiment sheds light on the radically different cognitive patterns between two core moral judgements: We require much more of actions and agents to judge them praiseworthy than to judge them blameworthy. The Sentimental Rules account can explain the judgement in harm scenario. In the harm scenario, it is possible that our care mechanism is activated because the environment is harmed due to the chairman’s decision, leading the participants to judge that the chairman is blame-worthy. However, it cannot explain why we apply different criteria in judging actions that to be morally praiseworthy. For example, participants make stronger demands of agents’ intentions and their relationship to the consequences of agents’ actions in order to judge agents praiseworthy than to judge them blameworthy. In the help condition, most participants believe that the chairman does not have the intention to bring about the side effect of helping the environment, thus he deserves little praise about what he has done. In this scenario, our norms for prohibiting harm are not activated, and there is no suffering of others that will trigger our affective response. But the participants are clearly making moral evaluation: judging whether the chairman is praise-worthy or not.

Let us consider another morally praiseworthy scenario: during the Covid-19 pandemic, we praised highly our front-line workers, especially doctors and other medical workers who risked their lives to save others’. In this scenario, from a perspective of a healthy citizen, we acknowledge three events/actions: vulnerable people are suffering from the horrible disease, medical workers are saving patients’ lives, and NHS workers are risking their own lives by working in the front line. If we only acknowledge the first two facts, then sentimental rules might be able to explain them. However, we clearly also recognise the third fact and NHS workers’ risk and sacrifice. This means that we also recognise that when NHS workers are helping others, they are harming themselves by intentionally putting themselves in a certain degree of danger. However, it would be bizarre to say that their bravery is morally wrong because it causes harm to them. Quite the opposite, we think more highly of medical workers because they are brave and fearless. There must therefore be other evaluative processes involved in judging an action as morally praise-worthy. This also reveals another problem for the sentimental rules account in judgements of moral goodness: it is far from clear that the valence of affective emotional response determines the outcome of moral judgement. In other words, some negative emotional responses can constitute judgements of morally good as well. The death of Dr Li Wenliang has provoked tremendous grief and the anger of billions of Chinese people. Li was the first person to warn that there might be an outbreak of a SARS-like virus. However, he was admonished because of showing his worry on his own social media. Unfortunately, the outbreak of COVID-19 happened later on, and he got infected and sadly died from the disease. He was dubbed as a whistleblower. People consider him as a national hero who tride to reveal the truth of the fact and save others’ lives. In this sad scenario, there is no positive emotion involved in people’s judging Li’s action and personality. Instead, we feel angry about how the police and the authority treat Li unfairly, we feel sad about his death, and we feel grief that we should have avoided the sad outbreak of COVID-19. These negative affective responses altogether make us praise highly of Li’s behaviour and treat him as a hero after his death. If he was treated properly by the authority, and the authority had paid attention to the warning, it is possible he would not receive this high level of praise among the public, as he was just doing his job as a professional doctor. He might be praised to some certain degree, but he would not be considered as a national hero.

As aforementioned, my ongoing argument against Nichols’s view is based on implications from both scientific experiments and everyday life events. I think evidence from experiments and everyday life share the same weight of persuasive powers. Scientific results from psychological experiments is a great enterprise of experimental philosophy, yet at the end of the day, morality is manifested in our daily life, rather than in the lab. We can observe how people actually make moral judgements in addition to putting them into an experimental environment. As far as I am concerned, the link between affective response and moral judgement is not as direct as the sentimental rules account has sketched. In order to make the sentimental rules account more plausible, Nichols at least needs to explain what mechanism is involved in distinguishing good/bad moral judgement, because as can been seen from my analysis above, our affective responses are not sufficient to moral judgement , and it is not enough to distinguish between morally good and morally bad judgement.

1. Implications

So far, I have outlined Nichols’s own work on moral sentimentalism and how he mounts a defence of his position through various appeals to empirical data. I have argued that the empirical evidence used to justify Nichols’s sentimental rules account does not directly support his empirical claim regarding moral judgement. There is a wealth of competing evidence and thought experiments that undermine the ability of the sentimental rules account to explain moral judgement, particularly cases where sentimental rules are triggered in non-moral scenarios which result in non-normative judgements. In addition, moral judgements are underpinned by different cognitive mechanisms, as there are at least two types of core moral judgement: one pertaining to moral permissibility and the other to moral impermissibility. In this final section, I will explore the relation between empirical work and philosophical research. Based on my consideration of Nichols’s work in relation to psychological experiments, it seems that although scientific discoveries have shed promising light on philosophical investigations of morality, there are still gaps between them.

The first lesson is that when we find a key element in a type of moral judgement, it cannot be naturally considered as the key mechanism behind all moral judgements. For Nichols, harm-based moral judgement is one of the core moral judgements. However, as mentioned above, we cannot derive a general cognitive mechanism from harm-based moral judgement because harm cues are not the only factor involved in moral judgement, nor is it the case that when our harm-prohibited mechanism is activated that it necessarily triggers moral judgement. When moral judgement is initiated the harm-prohibition mechanism sometimes remains idle, especially when we are making positive moral judgments.

Aside from the possibility that there might be two different mechanisms underlying judgements of moral goodness and moral wrongness, there are empirical studies indicating that different kinds of negative moral judgements can imply different neural mechanisms as well. For example, Parkinson et al. (2011) question whether ‘moral judgements’ comprise a homogenous class of behaviour and should therefore be studied scientifically as a unified category. Their experiment was conducted using fMRI to examine the neural structures of three different types of moral judgements: judgments pertaining to physical harm, dishonesty and sexual disgust. The results show that the regions of brain activity and the types of neural activity underlying these three types of moral judgements are discernibly different. Their findings reveal that dishonesty is associated with the brain area mainly responsible for mentalising, disgust is associated with affective processing and judgements pertaining to harmful moral transgression are associated with regions of the brain responsible for action understanding. The results bear significant implications for the traditional philosophical discourse on moral judgement, which, at its heart, makes a clear distinction between our emotions and rationality. The sentiment/rationality dichotomy is the core framework for debates in moral psychology: either you claim that emotions are essential, or *vice versa*. Hence, the challenge presented by the findings is that it seems both emotional and rational faculties can affect moral judgements: the act of judging an action as being dishonest involved the feeling of disgust as well as the capacity to reason that the action involves physical harm. On this picture it seems possible that different mechanisms may be essential to different kinds of moral judgements, thus it would seem too coarse to say *either* emotion or reason is essential to moral judgement. So, while the Sentimental Rules account might explain some moral judgements, it does not explain the mechanism underlying all *core* moral judgements as Nichols claims.

The second lesson is that conducting experiments on moral/conventional tasks is not a reliable way of testing participants’ ability to make moral judgments. For Nichols, to see whether participants are able to distinguish between moral and conventional judgements is a reliable way of determining whether they have moral concepts and whether they utilise such concepts in making moral judgements. However, there are studies showing that the distinction between moral and conventional judgements are very unclear, and whether agents consider an action as morally wrong or conventionally wrong heavily depends on both the background of the participants and the context of the given moral vignette. It remains an active debate as to whether there is a difference between moral and conventional judgements (Kelly et al. 2007; Quintelier 2010). We seem to have an intuitive grasp of the difference between moral and conventional rules, and individuals tend to be able to identify prototypical cases of morally wrong scenarios. However, in the philosophical literature there is no consensus about the emergence of morality and moral understanding.

Psychologist Elliot Turiel (1983) has provided three core features of moral judgements that purport to distinguish moral judgement from conventional judgement. According to Turiel, moral judgements are distinguished from conventional judgments according to their universality, authority-independence and about their incorporation of justice concepts (Turiel 1983; Stich 2018). These three features have been applied to test the moral understanding of agents in various psychological experiments and Nichols accepts these three features as the essential properties for moral judgements. However, some recent findings have challenged this claim. An experiment conducted by Kelly et al. (Kelly et al. 2007) sought to investigate whether harm-based judgement is actually more universal and independent of authority and thus qualifies as distinct from conventional judgement. Contray to Turiel’s principles, their findings suggest that (1) moral judgements are in fact authority-dependent and not universal; and, (2) harm is not the key feature in distinguishing moral judgement from conventional judgement. The experiment was conducted by presenting participants with a single historical scenario in two different contexts, in order to gauge the variation of participant responses in relation to the different contexts. The scenario in question was one of whipping sailors according to two different historical backgrounds (Kelly et al. 2007, 123–24):

Background 1: *Three hundred years ago*, whipping was a common practice in most navies and on cargo ships. There were no laws against it, and almost everyone thought that whipping was an appropriate way to discipline sailors who disobeyed orders or were drunk on duty. Mr. Williams was an officer on a cargo ship 300 years ago. One night, while at sea, he found a sailor drunk at a time when the sailor should have been on watch. After the sailor sobered up, Williams punished the sailor by giving him 5 lashes with a whip.

Background 2: Mr. Adams is an officer on a large modern American cargo ship in *2004*. One night, while at sea, he finds a sailor drunk at a time when the sailor should have been monitoring the radar screen. After the sailor sobers up, Adams punishes the sailor by giving him 5 lashes with a whip.

When participants are asked if it is okay for Mr. Williams and Mr. Adams to whip the sailor, majority of them answered it is okay for Mr. Williams but it is not okay for Mr. Adams to whip the sailor. This shows that whipping is morally impermissible nowadays but it is morally permissible 300 years ago. It further suggests that moral rules regarding harm-prohibition are not generalised and universal. As a result, it is possible that Blair’s experiments cannot reveal whether psychopaths cannot make moral judgement based on their antisocial behaviours, as the task of sorting moral and conventional judgments is not a good way of testing agents’ moral capacities at all.

Thirdly, the literature of moral psychology should give more attention to the mechanism of how we make positive morally judgements. While there is a large literature on altruistic behaviours, altruistic behaviours are not identical judgements that actions are morally good. We can be motivated to act altruistically more frequently than to judge an action is morally good: we have a higher standard or complicated standards for morally good actions. They are also a kind of core moral judgements, as it leads to how we will value virtue in our communities, and what actions we would like to promote in order to have a better society. It is possible that different neural mechanisms underlie different core moral judgements.

In a nutshell, there is no direct link between recent empirical findings and meta-ethical commitments. It is a complicated issue. So far, various psychological experiments on morality, moral development, and moral evaluation are conducted. They can play a crucial part in better understanding how we make moral judgements, and studying morality from a naturalistic approach. However, one psychological finding might only be able to reveal a factor of some moral judgments, yet not provide a general account of morality. In the case of Nichols’s argumentations, psychological experiments on psychopathy and autism cannot support that sentiments are essential to moral judgements.

## Chapter 3 Wrap Up: The Role of Emotions in Moral Judgements

In my previous two chapters, I have examined two influential accounts which argue for the essential role of emotions in moral judgements. This is a wrap-up chapter further investigate the role of emotions in moral system. The main claim is that current literatures can confirm a causal link between emotions and our normative system, however, it cannot further justify the sentimentalists’ claim that emotions are essential to normative systems. In my previous chapter, I have targeted some of the popular sentimentalists’ accounts. In this chapter I will further contend that sentimentalists are wrong by investigating the literature on disgusts and normative judgements.

1. Literature Review on Disgust and Normative Judgements

It has been a long tradition in moral philosophy to use reasoning/sentiment distinction as a start-up theoretical framework to think about the nature of human moral faculty: either you are a rationalist, arguing that moral judgements are grounded in rationalised norm and rules (like Kant), or if you think moral judgements are affective expressions of ones’ desire or feelings, you are a sentimentalist (like David Hume). This dividing way of thinking normative system is further reinforced as the topic gradually attracted scholars from other disciplines, especially when psychologists also participate into the debate of the nature of normative system since 1970s. Their participations provide more empirical data in addition to conceptual analysis to think about the nature of moral judgements.

At the beginning of the psychological (empirical) research on normative system and moral development, rationalism was the favourite view. Based on the rationalist hypothesis that moral judgements are the results from deliberate reasoning, psychologist Lawrence Kohlberg (1986)provides a 6-stage moral development account, a structure aiming to describe how humans develop moral judgements, and how the maturity of moral cognition is directly related to one’s reasoning and cognitive capacity.

Kohlberg’s theory and his rationalist claim on moral cognition has been dominant in moral psychology for decades until early 21st century, when Jonathan Haidt dropped a bombshell in the discipline with his highly influential and his up-to-date most cited paper ‘The emotional dog and its rational tail: A social intuitionist approach to moral judgment’ (Haidt 2001). The paper does not come out of blue: it is based on work on moral and cultural development done by cultural psychologist and anthropologist Richard Shweder, and Haidt his own early research on disgusts in relation to social evaluative judgements back in 1990s (Way, Shweder, and Haidt 1993; Haidt, Koller, and Dias 1993a). The main conclusion they initially drew was that cultural norms and normative judgements are substantially shaped by emotions. Meanwhile, when we are making normative judgements, we are expressing our emotions rather than rationally evaluating whether the actions are right or wrong. The main evidence they rely on was that people tend to judge some harmless acts morally wrong (for example, cleaning your toilet with national flags, eating your dead pet dog). Most of the time when people are making such moral evaluations, they failed to provide a justified reason. Instead, they seem to merely express their sentiments towards the acts (‘It is just disgusting!’), and these phenomena happen cross-culturally. In his 2001 paper, Haidt further provides a model, namely Social Intuitionist Model, to theorise their findings. According to Social Intuitionist Model (SIM), moral judgements are made via an affective and intuitive process effortlessly and quickly, while reasons are given after the judgements have been made. Although there are some exceptions where people initiate reasoning to make moral judgements, most of the time people do not do that. The typical case he referred to in this 2001 paper is sibling having protected sex.

SIM had a profound impact on the development of literature of moral psychology in 21st Century. Not only that the favourite position shifts from Kohlberg’s rationalism to sentimentalism, but also that disgusts become the most favourite subject for empirical researchers to study affects and decision-making(See Lerner et al. 2015 for review ). Empirical researchers are particularly interested in two issues: one is how unrelated *incidental disgusts* could profoundly shape our moral judgements: whether when participants are experiencing the emotions of disgust, their judgements are harsher than the controlled group. And in relation to the incidental disgusts, researchers also further explore whether individuals’ biological sensitivity to disgusts would play a role in amplifying biases (for example: racism, sexism and homophobias), and influencing moral judgements when disgust is induced. For this section, I will mainly focus on the former case.

In those which focus on studying induced disgust and normative judgement directly, there are four main strategies of inducing disgust in the experiment: exposure to a noxious odour (Schnall, Benton, and Harvey 2008; Białek et al. 2021), exposure to disgusting images (Cameron, Payne, and Doris 2013; Jylkkä, Härkönen, and Hyönä 2020; Sanyal, McAuliffe, and Curry 2021), tasting a bitter drink (Eskine, Kacinik, and Prinz 2011; Ghelfi et al. 2020),recalling or writing about disgusting experience(Johnson et al. 2016; Schnall, Benton, and Harvey 2008). In addition, Wheatley and Haidt (Wheatley and Haidt 2005) manipulate agents’ emotional states and attribution by connecting neutral words (either ‘take’ or ‘often’) to disgust when participants were in hypnotic state. After they were brought out of the hypnotic state, they would be given six vignettes which describe different moral transgressions: second cousins who had a sexual relationship (incest), a man eating his already dead dog, a congressman taking bribes (bribery), an ambulance-chasing lawyers, shoplifting, a student stealing library books. For the six vignettes, there are two versions while the contents are identical but one with the word ‘take’ while the other with the word ‘often’.

The results and conclusions are mixed. In the case of induced disgust on normative judgements, there are researchers claiming that their experiments show robust effects of disgusts on normative judgements(Schnall, Benton, and Harvey 2008; Eskine, Kacinik, and Prinz 2011; Wheatley and Haidt 2005). For example, in the scenario of olfactory disgusts, in their first experiment, Schnall and colleagues (2008)found out that people who are exposed to disgusting smell tend to make harsher judgements compared to controlled group. Meanwhile, in their other 3 experiments introduced in their paper, they found out that if the participants are highly sensitive to the feelings of disgust, they would more incline to more severe moral judgements compared to controlled group. In the scenario of beverage disgust, Eskine and colleagues (2011) found out that people who tasted bitter drinks tend to judge harsher. However, later research (Ghelfi et al. 2020; Białek et al. 2021), which recruited more participants in their study and attempted to reproduce the claimed effects of disgust, have fail to reach the same results. Regardless of the dispute, perhaps a vast majority of researchers happily agree that emotions such as disgust can causally influence moral judgements in a negative way sometimes. However, researchers such as Schnall, Wheatley and Haidt seem to have more ambitious goal. Based on their experiments, they tend to draw a much stronger claim: emotions do not simply colour or amplify moral judgements, rather, emotions play an essential role in normative judgements. For example, Schnall and her colleagues concluded in their paper that their experimental data indicate that gut feelings can fundamentally shape moral judgements(Schnall et al. 2008, 1100), Prinz (Prinz 2006; 2007) further draws a philosophical conclusion that moral judgements are essentially constituted by emotions. This line of thought and position is known as sentimentalism.

When ‘sentimentalist’ is becoming a popular label for moral psychologists, some other researchers also challenge sentimentalists’ claim, and their empirical findings as well. In general, they deny that empirical evidence is robust enough to show that gut feelings can shape our moral judgements in a profound way (May 2014). There are 2 arguments given by sentimentalists’ opponents: one is to refer to the fact that some of the results that show robust effects of disgusts on normative judgements cannot be replicated, the other is to cast doubt on publication biases in disgust research.

Let us talk about the first reasons with an illustration of studies of disgust drinks. In the case of inducing beverage disgust, Eskine, Kacinik and Prinz (2011) split 57 participants into 3 groups: bitter drinks (Swedish Bitters), sweet drinks and controlled group(water), each group of participants were given the same vignettes as that in Wheatly and Haidt (2005). They found out that while judgements in the control and sweet conditions did not different significantly, participants’ moral judgements in disgust conditions are significantly harsher. However, Ghelfi and collaborators (2020) failed to replicate the results when they conducted a large-scale (N= 1137) study of beverage disgust and moral judgements.

Regards to the second reason, in their metanalysis of literature of disgust, Landy and Goodwin (2015)revisited relevant work on disgust and moral judgements, arguing that disgust only shows robust effect in olfactory disgusts. In the case of beverage disgust, for example, later experiments cannot replicate the results that bitter drinks make moral judgements harsher. Meanwhile, Landy and Goodwin also considered the data from unpublished work. After taking unpublished work into consideration, they claim that in general, the results cannot show that disgusts can fundamentally affect moral judgements.

Schnall and her collaborators replied to Landy and Goodwin saying that the literature did support their 2008 conclusions(Schnall et al. 2015). There are two explanations about why sometimes effects of disgusts on moral judgements are not robust. One is that other moderator variables are interfered. For example, in the scenario of imagining disgusting scene, other cognitive processes such as imagination, memory and individual differences in disgust sensitivity, are involved and interfered. At the same time, Schnall and colleagues disputed publication biases objection, claiming that unpublished data should not be considered, as ‘relevant methodological details are not reported—the resulting effect size estimates are highly unreliable’ (Schnall et al. 2015, 538)

1. Empirical Results Cannot Justify Moral Sentimentalism

As can be seen from the analysis above, in the literature of disgust and normative judgements, there are empirical studies suggesting that moral judgements can be affected by irrational and intuitive processes. According to the findings, some researchers have speculated that emotions are essential to moral judgements. At the same time, this claim is also challenged by other researchers, and it is also the case that some of the results cannot be replicated. I argue that based on the current findings and experiments, we can infer emotions can influence our normative judgements in certain level, and there is(are) a causal relationship(s) between affective states and normative systems. However, current data *cannot* conclude that emotions are essential to moral judgements, or moral judgements are grounded in emotions as typical sentimentalists have claimed.

First of all, there is an explanatory gap between the findings that emotions can influence moral judgements and the metaethical claim that moral judgements are emotional in nature (to make a moral judgement is to express a sentiment of approval or disapproval). Admittedly, emotions can influence moral judgements, just like emotions can influence and manipulate other human cognitive faculties. However, it cannot further prove that emotions are the essential and fundamental parts of those cognitive faculties, including our normative systems. Consider our language system. The interference of emotions of course influences our expressions, however it cannot further generalise that linguistic expressions are emotional in nature. For example, let us consider an expression: ‘I am writing my third chapter’. This is a statement of my intention to finish writing a chapter. My emotional states will undoubtedly affect my mental states, further affecting the overall meaning of my expression: suppose I am in the office writing my chapter and getting stuck, while my PhD fellow PhD student James just walked into the office. He asked me: ‘what are you working on? You look a bit panic!’. I might reply impatiently: ‘I am writing my third chapter! I don’t really have time to talk now!’. Clearly, in addition to telling James that I am working on my chapter, I am also expressing my agitations about writing the third chapter. Meanwhile, for the same linguistic expression, I might also intentionally use emotions and sentiments to strengthen my point: suppose my supervisor Gerrard sent me an email checking on my progress: ‘hey how is everything going?’, and I replied: ‘Hi professor, I am currently writing my third chapter, and I am very confident that this is going to be a great chapter!’. In this scenario, I am deliberately adding positive emotions into my expressions, intending to show that I am not only writing my chapter, but also that I am very optimistic and feeling great about the chapter. At the first look, it seems that emotions are so fundamental to my expression: the change of my mood can completely change the valence of my expressions. I might convey different sentiments while expressing the exact same thing. But it doesn’t mean that your expression of certain thing is based on your emotions, and emotions need to be present in order for me to articulate my basic. For ‘I am writing my third chapter’ expression, emotions do not need to present in order to reply to both James and Gerrard: what is essential in my expression is that I am writing my chapter, and emotions are bonus. Without the articulations of the act (writing my chapter), neither James nor Gerrard would not know what I am doing. Languages have various functions: articulations, persuasions or expressing certain sentiments. However, it is bizarre to contend that expressing sentiments are essential to language: in order to express ourselves, we don’t need to express emotions the whole time, or have a disposition to have affective states. This is not to say that emotions are essential to language expression (also there are expressions that are purely for expressing sentiments). But for language system, other capacities seem much more important: the memory capacity of vocabulary and grammar, the reasoning capacity of detecting environmental cues, the implementing and inferencing capacity which is manifested in articulating myself, and so on. This is also the case in morality and normative cognition. In order to judge something wrong and bad, emotions do not need to be there. Most of time, emotions do add some spice to our judgements: it can make our moral judgements harsher because we are angry, or it can make our moral judgements more lenient because we are moved. But as far as I am concerned, to say emotions are essential to moral judgements is a big leap from our current data and daily-life practices.

Secondly, it is questionable whether participants’ response in lab setups is coherent and consistent with general everyday normative judgements. First of all, participants might be able to guess the purpose of studies, which will unintentionally alter participants’ initial answers. For example, in Schnall and colleagues’ 2008 experiments, they explicitly asked participants how disgusted they currently felt; whether they were consciously aware of any unpleasant smell while they were answering the moral judgements, if so how much the smells bother them. It is highly possible that participants have guessed the purpose of the study even though they have claimed that whoever guessed the experimental hypotheses have been excluded. Second, the vignettes and questionnaires are questionable, as they are mainly simplified versions of moral issues, and it might result in oversimplified moral judgements process. One of the main strategies for psychologists to investigate issues in moral psychology is via providing participants with stories which involve a certain level of moral transgressions, and they tend to be very intuitive and familiar to participants. It is possible that these moral transgressions are too familiar because we have already adopted relevant norms, the experiment setup further eliminate other necessary and essential process that might be involved in moral judgements. In a lab setup, normally each illustrated moral transgression is clear and direct: it is about one and only one moral transgression. Thus, when participants are presented with respective scenarios, they can quickly detect the moral cues from the event and make a moral judgement without efforts. This does not suggest that it is the most case that we can arrive at quick moral judgements most of the time. We encounter different moral cases in a daily basis, they are much more complex than what the vignette has illustrated, and they are not necessarily derived from a simpler version of moral transgression, especially when it comes to moral judgements about new issues. For example, when we judge whether the UK should be more open to refugees, whether we should stop consuming meat, and so on. These judgements are not merely about responding to one moral violation, they are about deliberately moral reasoning: is choosing fairness going to harm local people? Is sacrificing immigrants worthwhile for security? And so on. As a result, it is possible that when we are considering simpler situation, we react faster. It is questionable whether in social science, simpler questions and vignettes can provide understanding of how people solve difficult social and moral issues.

The third worry is also about experimental setup and empirical results. Even if it is the case that most of the time, we only deal with simple moral task, and it is possible that we morally judge things quickly and intuitively, but it does not further suggest that this quick and effortless process is mainly irrational sentiment. Although studies show effects on influencing the seriousness of moral judgements, studies also show that other cognitive faculty—for example, attentions and memories – can moderate the effects that disgusts have brought. Regards to this objection, Schnall and colleagues have responded that when researchers exclude other moderator variables from the context, disgusts do show robust effects on moral judgements (Schnall et al 2015, 257). However, we need to further question: what if other moderator variables are more important than disgusts? What if other cognitive capacities, such as attention and memory, show more robust effect on moral judgements? Few theorists deny that emotions have motivational power (this is one of the featured properties of affective system in general). But it is far from conclusive that we make most of our moral decisions based on how we feel, and other cognitive components are not as essential as emotions. For example, a crying baby with a dirty diaper is disgusting: they cry unreasonably loud and they stink. When their parents are changing the disgusting diaper, it is highly impossible that they will judge their baby morally. Why do we assume that parents would not judge their new-borns morally even if their babies can be annoying and frustrating sometimes (or most of the time)? I speculate that when we are making moral judgements, we also initiate non-emotional cognitive processes: we presuppose that babies cannot control their behaviours, they lack certain agential power, so we do not think babies should take the moral responsibility of being irrationally annoying and messy. These presuppositions require cognitive capacities such as mindreading, expectation and inferencing, which are not necessarily conscious and affective. In this crying dirty baby case, other cognitive factors undoubtedly overweigh the negative emotional impacts when we are dealing with this scenario. Now Let us consider a real moral case. Incest is a typical story used in different experiments: in the given story plot, two siblings (or close cousins) enjoyed sex without causing any further moral issues (no second time, no pregnant, and no one knows). It is understandable that researchers want to simplify the scenarios, cut off the unnecessary story branches, and let participants to directly initiate their moral judgements. However, in the reconstructed incest case, the norm ‘incest is wrong’ has already existed in our normative database before the experiments, and what the experiments did was to make it explicit to participants without considering other factors that might influence moral judgements. What participants really need to do psychologically, is to quickly initiate the moral judgement that incest is wrong. In everyday life experience, even for the incest case, other contexts will heavily influence moral judgements: whether the local community which the protagonists are in allow incest? Whether they are really committed to one-time off? Whether they are 100% sure that they will not have kids? And so on. These questions are normally taken into considerations explicitly or implicitly when people are making moral judgements about incest case, which suggest that apart from saying right or wrong about an action, in order to make a judgement, other factors are inevitably involved, and they also fundamentally influence moral judgements. Accordingly, the reconstructed and oversimplified incest case is peculiar in the way that it is somehow very counterintuitive: although researchers have specified that no harm would be caused by the given incest case, participants might have automatically taken the potential harm into considerations based on their previous social and norm learning. Normally when individuals are making a moral judgement, they need to detect moral cues from the environments and motivate response such as punishment or expectation. These are all part of the process of moral judgements, and they involve other non-emotion factor such as perception(Francis et al. 2017), memory(Ting, He, and Baillargeon 2019), inference (Gweon 2021)and attention(Cameron, Harris, and Payne 2016). These non-emotional cognitive factors are fundamental to moral judgements, and studies on disgusts in lab cannot rule out this proposal. Quite opposite, studies have shown that humans are able to make moral judgement based on their expectations of protagonists’ behaviour at the early stage. Ting and Baillargeon (2021) have shown that 2-year-old children draw a negative inference from a wrongdoer’s moral violation, further expecting that the wrongdoer is likely to do something wrong in other contexts. This has suggested that it is unrealistic for researchers to assume that as long as the given scenarios are not harmful, the scenarios appear harmless to participants. At the same time, it is also unrealistic for researchers to cut off other cognitive processes and just investigate the influence of emotions of moral judgements, further directly conclude that emotions are essential to moral judgements.

Last but not the least, it is troublesome for sentimentalists that several attempts to replicate the robust effects of disgust on moral judgements have failed, which have shaken the empirical foundation of sentimentalism. As I have illustrated previously, in a larger-scale study on disgusting beverage drinks and moral judgements, Ghelfi and collaborators (2020) failed to replicate the initial result reported by Eskine and colleagues (2011). Apart from examples of beverage drinks, similar problems also happen in in their first experiment, Schnall and colleagues investigate whether induced disgusting smells influence moral judgements. The moral transgressions provided are marriage between first cousins, sex between first cousins, driving rather than walking to work and releasing a morally controversial film. The results have shown that when participants are exposed to strong disgusting smell, they tend to make harsher moral judgements compared to controlled group. It is worth being noted that the sample size is 120, and all the participants are all from Stanford University, US. In another experiment about induced disgusting smell and moral judgements, Białek and colleagues (2021) investigated whether incidental disgust influences moral judgements in both moral transgressions and moral dilemmas. They initially recruited 142 participants from Maria Curie-Skłodowska University, Poland, and presented them with 20 moral problems and 10 moral dilemmas vignettes. This experiment found no robust effects of disgusts on neither type of moral scenarios. In their second experiment, without changing the major setup and vignettes (they deleted low-odour condition), they recruited more participants (n=248) and they still fail to replicate the results that Schnall and her collaborators have found. The inconsistency of the findings requires further interpretation: it is possible that sample size matters. When two experimental results are inconsistent with each other, the one with larger sample are more trustworthy. It is also possible that the role of disgusts on moral judgements are different across cultures. Maybe in some social communities, disgust is applied in moral development at the early age, while in some other countries, there are different strategies to teach moral values. Regardless, we cannot conclude that disgusts can fundamentally shape our moral judgements, and that emotions are essential to moral judgements.

# 

# Part 2: Moral Competence and Moral Learning

In part 1, I have explored the role of emotions in moral judgements. By referring to empirical studies, I have shown that emotions influence moral judgements significantly: the linguistic form of moral judgements normally contain emotional elements to express our moral attitudes directly. At the same time, emotions can also influence our moral judgements in an unwanted way by amplifying unwanted biases and heuristics. I argued that current empirical evidence cannot prove that sentimentalists are correct about moral judgements, especially the claim that sentiment is the ultimate source of moral judgements.

But the purpose of part 1 does not end here. In addition to my rejection of moral sentimentalism, we can reasonably see that moral rationalists will face similar problems as well. No rationalists would deny that emotions are crucial to moral cognitions, but they will also face the similar question: it is impossible to pinpoint a rational system that is emotion-immune, which grounds moral cognitions? Hence, it is not surprising that some researchers have proposed that the sentimentalism/rationalism dichotomy is not helpful for understanding the psychological mechanism of moral cognition (May 2018).

If we give up on the framework of sentimentalism/rationalism, we need another new foundation that begins to sketch out a psychological mechanism. An alternative way is to ground our moral capacity in *evolutionary psychology* (Tooby and Cosmides 2005; Curry et al. 2021)*.* Evolutionary psychology is a theoretical approach which tries to understand human psychology through the lens of evolutionary theory, in the sense that how our mind represents the world is fundamentally shaped by the process of evolution (Downes 2021). Following this line of thought, nearly all our cognitive capacities emerge driven by evolutionary processes. In other words, our mind consists of different modules that are for different cognitive tasks (Fodor 1983; Stich 2006; Carruthers 2006), including our moral cognition (Heyes 2018; 2021; Kumar and Campbell 2022).

Once we accept the claim that moral cognition is made possible because of evolutionary drive, the next question would be: what is the innate architecture for moral cognition? In this part, I will focus on two models. The first model has been influenced by research on human linguistics, especially Noam Chomsky’s linguistic theory, which argues that humans are equipped with a domain-specific mechanism—Universal Grammar—to detect linguistic cues in the environment and acquire linguistic knowledge. Universal Grammar is a result of evolutionary processes. Similarly, some researchers argue that we also have a domain-specific mechanism—Universal Moral Grammar—to deal with moral cognition. Universal Moral Grammar, which is a result of evolution, guarantees our competence for moral cognition. I will evaluate this model in chapter 4. The second model is called cooperation-based moral theory, which argues that moral cognition is for facilitating cooperation in humans’ social lives. And the need for cooperation is also a result of evolutionary processes: we need to cooperate with others for survival and reproduction.

## Chapter 4 Do We Have an Innate Universal Moral Grammar

As promised, this chapter will focus on the proposal which argues that moral knowledge implies a domain-specific psychological module for moral cognition. This view is known as moral nativism or moral Chomskyanism. Proponents of moral Chomskyanism argue that since the process of moral learning is very complex and intricate, and it is impossible for children to learn from the environment based on their cognitive status. However, human children are able to acquire and internalise moral norms. Hence, there must be a richly structured moral faculty (Hauser 2006; Mikhail 2007; Dwyer, Huebner, and Hauser 2010). Moreover, what makes this view distinct from other moral nativists’ views is that they have been influenced from Chomsky’s model on linguistics, arguing that the structure of moral knowledge is similar to that of language.

In this chapter, I will tackle this position and argue that moral Chomskyan’s argument—linguistic analogy—fails to pinpoint an innate and domain-specific architecture for moral cognition (if there is a such thing). The first section of this chapter will be a general introduction to the nativism and empiricism debate, and the arguments that humans’ cognitive performance in linguistics implies a linguistic competence, which is predetermined by an innate and domain-specific psychological faculty. There are two arguments for this proposal: the poverty of stimulus arguments and universal grammar (Chomsky 1976; 1980; 1986). I will also unpack these two important arguments in section 1.

After that, in section 2, I will illustrate the moral Chomskyan’s arguments: linguistic-analogy and universal moral grammar. According to moral Chomskyans, just as the capacity required for linguistic knowledge is rich and complex, the cognitive capacity required for moral cognition is also structurally complex. At the same time, we excel at moral representation in a fast, efficient and rapid way. This implies that our moral cognition should be elicited by an inherited structure (Mikhail 2011). I will then evaluate their arguments in section 3, and argue that the moral Chomskyan’s linguistic analogy strategy doesn’t work. This is because linguistic capacity and moral capacity are fundamentally distinct. The cognitive difference is manifested in four ways: the learning strategy, the internalising process, the interaction with other psychological states/events/functions, and lastly, the underlying biological premise. These differences should be sufficient to conclude that the similarity between language and morality is superficial (Sripada 2008). This means we could hit two birds with one stone, that linguistic analogy will ultimately fail and there is no Universal Moral Grammar. The fourth and final part of this chapter will be a conclusion of my arguments.

* + - 1. Setting the Stage: Nativism vs Empiricism Debates, Linguistic Capacity and Why It Matters to Morality

What is the nature of knowledge? How is knowledge possible? And how do we acquire knowledge? These are questions in epistemology and have been debated since the birth of philosophy. Cognitive scientists take different strategy to explore these core epistemic questions. Philosophers of cognitive science treat our mind as a *computation system* (not an actual computer), taking into consideration our psychological structure: if we treat our mind as a massive computation system, how is this computation system structured? How do the psychological structures of minds and mental states make knowledge possible to us? Does some knowledge imply the nature of the structure of our mind? To what extent is our mental structure innate?

When answering the above questions in the cognitive science battlefield, one rivalry is between nativism and empiricism. One of the ‘war zones’ is linguistics, and the early figure for each camp is Chomsky (1959) and Skinner (1957) respectively. It is worth emphasising that, nativism does not imply all the knowledge is innate, and neither does empiricism imply that our mind is a blank slate. Just as Laurence and Margolis have pointed out, the debate between nativism and empiricism is on ‘the quantity and richness of innate structure that they endorse’ (Laurence and Margolis 2001). Therefore, empiricists about knowledge do not hold that all the knowledge can be learned without endowing any innate psychological structures, and nativists do not hold that all the knowledge are a priori and experience is not important at all. The best way to reframe the debate would be: whether our mind consists of multiple modules that are predetermined? And what psychological systems are modular? In what follows, I will unpack the nativist’s view, especially their view on the nature of linguistic knowledge in detail.

* 1. A Cognitive System

Before discussing nativism from a cognitive science approach, we need to understand how cognitive scientists reframe the traditional epistemic question. As aforementioned, cognitive scientists treat the human mind as a computation system. But this claim is far from being practical for studying human minds scientifically: If we treat our mind as a computational system, and different psychological capacities as complex information-processing systems, from a practical perspective, how should we study and understand a psychological phenomenon in the lens of cognitive science? What kind of specific questions should we ask when we are tyring to build up a theory for understanding a psychological process?

Fortunately, there is an agreement on this. Cognitive scientist David Marr argues that there are three levels to understand a computation device: computation level, representation/algorithm level, and implementation level (Marr 1982, 23–25). At the same time, he also provides an example of supermarket cash register as an example. I will unpack them in detail.

For the computational level of understanding, we ask two main questions—questions of what and questions of why. What is the goal of this computation system (what is the task), and why is that the case? For example, the goal of a supermarket cash register is math, and usually it is doing addition. Hence, the cognitive task for a supermarket cash register is to excel at addition. At the computation level, we also need to explain why cash machine’s task is doing addition not other operations (for example, multiplication). As we can see, by understanding a system at the computation level, we should be able to understand the function of a system. On the other side, we should also know the limits and constraints of that system.

For the representation/algorithm level, we ask the question of how: how is the computation possible? Specifically, we can transform the how-question into two what-questions: what is the representation[[6]](#footnote-7) of input and the representation of output, and what kind of algorithm is implemented for the transformation between the representation of input and that of output? In the case of cash registers, the representations of input and output are the same: numbers. But in some other devices, the properties of the representation of input and output change. Just as Marr has presented, a Fourier transform’s input maybe the time domain but the output belongs to the frequency domain.

For the implementation level, we ask another how question: how can the representation and algorithm be realised physically? In the case of cash registers, for the same algorithm and representation, it is possible to be realised by different kinds of physical devices. According to Marr, to fully understand a computation system, all three levels of understanding are required (Marr 1982, 25). As we can see, by understanding the computation level, we are able to pinpoint the main function of a system. This is also the most abstractive level. By understanding the representation level, we should be able to look at the algorithm(s) and computational process(es) that make the function of that system possible. At this level of understanding, we should be aware that to achieve the same goal of cognitive function, it is possible that multiple algorithms and representations can achieve the same goal. That leads to the third level—the implementation level. It is also possible that different physical devices can realise the same algorithm.

* 1. Arguments from Nativism, Modularity of Mind and Innateness of Language

In general, nativism is the view that the mind has many complex domain-specific subsystems to differentiate different kinds of knowledge and concepts (Margolis and Laurence 2022). In other words, nativists not only commit to the claim that our mental structure contains *a lot of* innate materials, but also that our mind is highly differentiated: the many innate elements include lots of domain-specific cognitive systems. Each domain is differentiated by its cognitive functions or capacities. Therefore, for different cognitive capacities, abstract concepts, or principles of inference, they are realised by different modules.

In the following, I will unpack how nativists argue for the acquisition of language. According to Chomsky, the question on the nature of linguistic knowledge comes in three parts(Chomsky 1986):

1. What constitutes knowledge of language?
2. How is knowledge of language acquired?
3. How is knowledge of language put into use?

These three questions are the three fundamental questions when it comes to the study of the knowledge of language. As Mikhail has provided clear answers in his book (Mikhail 2011), according to Chomsky, in order to answer question (a) (what constitutes knowledge of language?), we need a theory of a particular *generative grammar* or a theory of *linguistic competence.* This theory should describe how a steady state of the mind of a person ‘knows’ or ‘cognises’ a particular natural language like English or Mandarin. To answer question (b) (how is knowledge of language acquired?), Chomsky proposes a model: *Universal Grammar,* which assume that there is a distinct subsystem in our mind that is devoted to language acquisition. This distinct subsystem is innately encoded in our genetic programme, which is unfolded by the modest trigger from the environment. Meanwhile, the interaction between UG and learners’ first-hand experience amounts to knowledge of a particular language. To answer question (c) (How is knowledge of language put into use?), Chomsky proposes that we need a theory of how knowledge of language is used to manifest actual expression in interpersonal interactions (Chomsky 1986).

As can be seen, these three fundamental questions target different aspects of the project of theorising linguistics. A successful theory of linguistic knowledge should be able to answer the three questions above. In addition, a successful theory of linguistic knowledge should also explain some unique phenomena in linguistic cognition.

The first phenomenon is the commonality and variability of human language. Human language exhibits a shocking commonality in terms of linguistic structure across human groups. For example, virtually all languages have nouns, verbs, adjectives for different functions. Nouns typically express objects, verbs typically describe actions, and adjectives typically describe states. In addition, all languages tend to have mechanisms for indicating tense, number and case. At the same time, language also demonstrates diversities across groups. Every language has its own unique vocabulary, syntax, phonetics and phonology, and semantics. A theory of linguistic knowledge should be able to accommodate this phenomenon.

The second phenomenon is the asymmetry between the linguistic information input and the linguistic performance output. When learning their first language, human children are not exposed to any explicit abstract sets of linguistic principles and grammars. However, human children are capable of mastering their first language learning.

* 1. Poverty of the Stimulus Argument and Universal Grammar

The most influential arguments from Chomsky for the innate domain-specific linguistic architecture is the Poverty-of-the-Stimulus argument (PoS). As aforementioned, nativism does not assert that experience is insignificant to knowledge acquisition, and that empiricism means that our brain is a blank slate. Hence, the exact definition of nativism is debatable. Accordingly, the Poverty of the Stimulus argument has been interpreted in different ways. Since this chapter does not directly concern Chomskyan linguistic theory, I will not delve into the debate here. Laurence and Margolis have helpfully provided a clear demonstration of the standard Poverty of Stimulus argument, and their interpretation has been widely accepted. Hence, I will just follow their construction (Laurence and Margolis 2001, 221):

1. An indefinite number of alternative sets of principles are consistent with the regularities found in the primary linguistic data.
2. The correct set of principles needn’t be (and typically isn’t) in any pretheoretic sense simpler or more natural than the alternatives.
3. The data that would be needed for choosing among these sets of principles are in many cases not the sort of data that are available to an empiricist learner in the child’s epistemic situation.
4. So if children were empiricist learners, they couldn’t reliably arrive at the correct grammar for their language.
5. Children do reliably arrive at the correct grammar for their language.
6. Therefore, children are not empiricist learners.

Several key terms need to be unpacked here. ‘Primary linguistic data’ refers to the language input that children receive when they are learning their first language. Premise 1) is the statement that there are countless languages out there in the world, and they have their own grammars and principles. However, they all share some similar patterns, and these patterns can be observed in the language input that a child received when they learn their first language. Premise 2) states that this set of principles is very complex and unintuitive. Premise 3) states that the set of linguistic principles are not the same kind of information children have when they started learning their first language. When children are learning their first language, parents do not teach them the abstract grammar. At the same time, it is questionable if children comprehend any of the abstract and complex grammar even if parents teach them the grammar. Both premise 2) and premise 3) yield to premise 4): children *cannot*, via empirical processes,learn the correct set of linguistic principles due to the complexity of the structure of linguistic principles and how it differs from the actual information children receive when they are learning their first language. However, children can acquire the correct grammar and competent at understanding and using their first language (premise 5). Hence there must be an innate psychological faculty that is able to detect linguistic cues, represent and internalise them, which makes children able to possess language.

In addition to this negative argument against empiricism, linguistic nativists such as Chomsky also argue that the sets of principles in language case are Universal Grammar (UG). This hypothesis argue that all human languages share a common structure, which is hardwired into the human brain. This common structure is called ‘Universal Grammar’. The main function of UG is that it consists of several constraints and principles that are specified for language. Hence, learners rely on UG to accurately detect linguistic cues from the environment, represent the input in a certain way. Hence, UG is the knowledge of linguistic competence, which is guaranteed by an inherited linguistic device by evolution. Hence, the existence of UG can explain both the commonality of the linguistic capacity across cultures, and the variety of different languages that are practiced by different groups.

It is worth making a clarification here: while nativists argue that the nature of our modular mind is innate, they do not infer that any supernatural being has designed the structure of our brain. Instead, they refer to *evolutionary psychology* to explain why our brain is wired in such way. For example, for those who endorse the theory of massive modularity of mind, they assert that the modularity mind is for adaptation and fitness which is driven by evolution (Robbins 2017): Just as human biological structure, the structure of human mind is a result of natural selection. In order to survive and reproduce, our human ancestors had to solve a number of recurrent adaptive problems (finding food, shelter, mates, etc.). Moreover, all the adaptive problems need to be solved more quickly, efficiently, and reliably. By comparison, a multi-modular system should be a better cognitive processing candidate than a non-modular one. Hence, natural selection would have favoured the evolution of a massively modular architecture. Therefore, human mind should be massively modular for different psychological tasks.

* + - 1. Linguistic Analogy for Moral Nativists

In section 1, I have illustrated linguistic nativism and nativists’ argument for the innateness of linguistic capacity. Chomsky and his followers have provided several important arguments to defend linguistic nativism. They have pinpointed two key features in human language: one is the commonality within diversity of human language, and the other is the shocking capacity children demonstrate when it comes to first-language acquisition. They have embedded these two phenomena into the Poverty-of-the-Stimulus argument to show that an innate domain-specific psychological mechanism is required. They name this psychological mechanism Universal Grammar.

Chomsky’s approach to the knowledge of language has been influential, not only in linguistic studies, but also in other studies. Just as other moral psychologists, a mission for moral nativists is to explain the paradox of moral norms: some moral norms are universally shared by human races, while some moral norms are local and shaped by cultures. Moral nativists try to answer the question of which innate structures are required for acquiring moral norms.

For example, Mikhail also unpacks the question of moral knowledge in three parts: the constitutions of moral knowledge, the acquisition of moral knowledge and the execution of moral knowledge (Mikhail 2011). Meanwhile, in several places (Mikhail 2007; Pinker and Fodor 2005; Dwyer 1999; Dwyer, Huebner, and Hauser 2010; Hauser 2006), moral nativists also argue that just like language, human moral performance entails an inherited moral competence resulting from evolution. In the following, I will demonstrate Mikhail’s linguistic analogy argument. Especially, I will focus on his argument for the complexity of moral codes manifested in representing trolley problems.

When making the linguistic analogy arguments, Mikhail focuses on three arguments, and later transformed them into three important premises for his overall linguistic analogy and universal moral grammar. The first is that our moral rules and principles are complex just like linguistic grammar; the second is that children are able to accurately acquire moral rules and principles without explicitly being taught so at a young age; and the last is that people excel at engaging in moral cognition in a fast, efficient and reliable way. The rest of the linguistic analogy argument is more or less similar to the Poverty of the Stimulus argument, so I will not repeat it here. In what follows, I will focus on how Mikhail justifies these three premises. If these three premises are proved, then the rest of Mikhail’s arguments would flow naturally similar to the original poverty of stimulus and universal grammar arguments.

These three premises are not self-justified, as we can immediately ask two questions related to these premises. First, how complex are the moral rules that we need to learn? This question is directly related to the first premise. Second, to what extent does moral knowledge imply a complex and richly structured moral faculty? This question is directly related to the second and the third premise.

Firstly, to demonstrate that moral rules are complex and require an inherited complex representation, Mikhail refers to the classic trolley problem in moral philosophy debates. Specifically, by manipulating the scenarios and the conditions, Mikhail found out that people will ascribe different moral obligations to the actors, and different levels of permissibility to whether the actors should save more people, which suggests that in the case of the trolley problem, rich and complex mental representation is manifested.

Initially, the trolley problem was a thought experiment raised by Phillipa Foot when discussing the problem of abortion and the principle of double effect (Foot 1967). The original trolley problem is simple: suppose that a driver of a running tram who can only steer from a narrow track to another, five men are working on one track and one man is working on another track, whether it is permissible for the driver to steer to the track with one man. Later on, Judith Jarvis Thomson (1976) discussed trolley problems when she investigated the difference between killing and letting die.

Based on the basic trolly problem scenario, Mikhail has provided six different trolley problem conditions to participants (Mikhail 2007, 149):

**Bystander:** Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. There is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die.

**Footbridge**: Ian is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. The heavy object is a man, standing next to Ian with his back turned. Ian can throw the man, killing him; or he can refrain from doing this, letting the five die.

**Loop track:** Ned is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. The heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die.

**Man-in-front:** Oscar is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. There is a man standing on the side track in front of the heavy object with his back turned. Oscar can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die.

**Dropman**: Victor is standing next to a switch, which he can throw, that will drop a heavy object into the path of the train, thereby preventing it from killing the men. The heavy object is a man, who is standing on a footbridge overlooking the tracks. Victor can throw the switch, killing him; or he can refrain from doing this, letting the five die.

﻿**Collapse bridge**: Walter is standing next to a switch, which he can throw, that will collapse a footbridge overlooking the tracks into the path of the train, thereby preventing it from killing the men. There is a man standing on the footbridge. Walter can throw the switch, killing him; or he can refrain from doing this, letting the five die.

These six trolley problems have shared the same overall features: all the protagonists are facing the same choice of whether they should throw a switch or a man in order to save five people’s lives while sacrificing one person. Secondly, neither scenario has an ideal ending, as no matter how the protagonists choose, there must be someone being killed. Regardless of these similarities, according to his study, participants have different moral judgements towards these actions. For example, participants judged that in the case of footbridge condition, it is morally impermissible to throw the man to save five people’s lives, while in the case of bystander condition, it is morally permissible for Hank to throw a switch to save five people’s lives even if one man would die in this scenario. This has shown that different moral rules are implemented by the moral representation of the actions, which result in different moral judgements. However, if we only analyse footbridge and bystander scenarios, other simpler accounts might already be able to solve the moral representation. For example, Greene (2001) has proposed a dual-process model of moral judgements, which argues that two kinds of psychological processes constitute moral judgements depending on whether the action is impersonal or personal. When the scenario is not personal, individuals will tend to initiate reflective reasoning to resolve the trolley dilemma. This reflective reasoning is controlled, slow and deliberate. When the scenario is personal, individuals would be driven by intuition to make moral judgements. This intuitive reasoning is fast, automatic and effortless. If Greene is correct, then moral representation is not complex, and it is possible that humans can acquire moral competence easily. To avoid this, Mikhail also discussed the other four scenarios. If our moral representation is as simple as the dual-process model suggests, then all impersonal scenarios should demonstrate similar moral judgements. However, this is not the case. For example, among all the six cases, only one case—the footbridge case—is a scenario which has a personal element (that Ian has to throw a man rather than throw a switch). If Greene is correct, moral judgements of the rest of cases should be similar. However, Mikhail’s studies show that they all demonstrate different patterns of moral judgements. The variety of moral judgements entail that people’s moral representation of each case differs. This means that moral cognitions are far more complex than Greene has suggested: people take into consideration actors’ intentions, the ends and means of the acts, as well as side effects when they are structuring the scenarios for moral consideration.

Mikhail further concludes that the results in the trolley problems studies suggest that a richly structured mental representation of actions influence moral judgements, and this is the reason why consequence is not the only consideration for moral evaluation. Hence, the study of trolley problems answers the first question: our moral representation is very complex, and there are far more moral rules and codes that we normally are not aware of. Now, here comes the second question: do these moral rules and moral codes entail a domain-specific moral faculty?

As aforementioned, Mikhail contends that the domain-specificity of morality can be manifested in two phenomena. The first is that children can acquire moral knowledge without being explicitly taught it. For this phenomenon, Mikhail refers to studies from developmental psychology to show that at early stage, children have acquired highly sophisticated and abstract moral concepts such as justice, and children are able to use motives and intentions as criteria for moral evaluation at 3 years old (Nelson 1980). Now let us turn to the second phenomenon: humans are able to entertain moral properties in a fast, efficient and reliable way. For this phenomenon, Mikhail does not provide a lot of evidence for it, he mainly took it for granted that moral judgements are intuition judgements, which take place very fast and effortlessly without lots of cognitive load.

Now we can reconstruct the arguments for the Universal Moral Grammar of moral representation[[7]](#footnote-8):

1. For the actions that brings out the same consequence, the moral representation may differ with according to the actors’ intentions and the action’s side effects.
2. Hence, moral representation entails a complex and richly structured cognitive system.
3. Children are able to engage in moral representation very well without being explicitly taught to do so.
4. Hence, there must be some innate moral representations.
5. Moral judgements are fast, intuitive and reliable.
6. Hence, moral representations are locally operated and dissociable to some extent.
7. Therefore, our moral faculty is an inherited domain-specific psychological architecture.
   * + 1. Evaluation

In this section, I will evaluate and assess Chomskyan moral nativists’ view. There are some advantages for moral Chomskyanism. The first advantage comes from its implication for moral epistemology. In moral epistemology, there has been an argument on the nature of moral knowledge: whether there is moral knowledge, and how moral knowledge is possible (Lenman 2007). In the camp of moral scepticism in moral epistemology specifically, there is a view which denies both the existence of moral properties and truth aptness of moral knowledge (eg., Olson 2010): moral properties are not a natural kind, and no one ever knows that any substantive moral belief is true (Sinnott-Armstrong 2006). Moral Chomskyanism is in a good position to refute this kind of moral scepticism. If moral Chomskyanism is right about moral cognition, we can treat moral knowledge as a natural kind, because our moral cognition is a psychological function that is realised by a domain-specific psychological mechanism. The psychological mechanism itself is a modular computation system that is operated following natural law and principles, hence its products—moral properties, belong to natural kinds as well.

This will lead to the second advantage of moral Chomskyanism: by grounding moral knowledge in a natural kind, we can study moral knowledge and moral cognition in scientific ways. What moral Chomskyanism has provided is a model of moral cognition at the computational level, and based on this, we can further move on to explore the psychological architecture of moral cognition from the algorithmic level. What this means is that we can further conduct scientific experiments or research to explore what neural structures, or specific psychological processes are involved in moral cognitions. Doing so we will get closer to deciphering the properties of universal moral grammar (Mikhail 2007).

This is all true only if moral Chomskyanism is right about the psychological architecture for moral judgements. But the view faces much more trouble than these potential advantages are worth.

There are two strands of argument that can be used to challenge moral Chomskyanism. The first one is to refute the modularity of mind or refute the Chomskyan linguistic proposal at the beginning. By refuting the theoretical foundation for universal moral grammar, the moral Chomskyan nativists’ argument would collapse. This might be a good approach: not every philosopher of cognitive science are persuaded by the view that our mind processes information and represents them in a modular way (P. S. Churchland 1987; P. M. Churchland 1988b; 1988a; P. S. Churchland and Grush 1999). Moreover, there are arguments against nativism and the Chomskyan view on linguistic knowledge (eg., Cowie 1999; Everett 2005), as well as the popular view that human thinking is structurally language-like (Fedorenko and Varley 2016).

But in my evaluation, I will not take this approach, because moral Chomskyans can still get away. They could respond by arguing that even if we do not have the innate linguistic acquisition device that Chomsky suggests, and even if most of our mental materials are not innate, it is possible that our moral cognition is domain-specific due to evolutionary demands, because morality is so fundamental to humans as social animals (Kelly and Setman 2021). Maybe it is the case that language is not domain-specific, but our moral cognition is. In our social lives, we need to be able to engage with moral cognition in a very fast, efficient and reliable way, which suggests that we need to be experts at moral cognitions with limited cognitive controls (Birch 2021). If we adopt an evolutionary psychology approach, then our moral cognition is possibly achieved by an innate domain-specific device. Hence, there is a way for moral Chomskyans to argue for universal moral grammar.

My focus is on moral cognition, and I do not intend to argue against massive modularity of mind as a whole. What I will try to do in this chapter, or even throughout the whole thesis, is just to reject a domain-specific moral faculty. Therefore, I will take the second approach in my evaluation. I will assume and adopt Carruthers’s view that our mind is massively modular, and that Chomsky is right about linguistic knowledge. What I will try to argue is that linguistics and morality do not share the similarity fundamentally: language and morality is not the same thing, therefore the linguistic analogy argument fails. Moral cognition and linguistics are different in three essential ways: the logic of learning strategy, mental representation, biological realisation. These three problems echo Marr’s three-level of understanding a cognitive system. While empirical evidence indicate that linguistic knowledge demonstrates domain-specificity in all three levels, and universal grammar exist, it is the opposite story for moral Chomskyanism.

In doing this, the non-modular feature of moral cognition will be revealed: it is centrally accessible and able to be changed by cognitive controls as it is cognitively sociable with other psychological functions. Meanwhile, it also suggests that the cognitive function of morality does not demonstrate the existence of universal moral grammar, which indicates complex rules and principles for moral norms acquisition. This will suffice to reject the domain-specific proposal and existence of universal moral grammar. Starting from this approach, we can better see that it will be a dead end to refer to linguistic concepts and theories to frame a theory of moral cognition.

* 1. Problem from Logic of Learning Strategy

The first difference between morality and linguistics is from the computation level, which is manifested in the different logics of learning strategy. This further reflects the different structure of these knowledges.

At the computational level, we ask the question of ‘what [is] the goal of that system, why is it appropriate, and what is the logic of the strategy by which it can be carried out’ (Marr 1982, 25). Chomskyan linguistics are able to provide a persuasive innate domain-specific argument for the logic of how the function of language acquisition be carried out. There are two reasons for adopting it: the first is from the poverty-of-stimulus argument that I have demonstrated in section 1, so I will just briefly state it here. Linguistic knowledge comprises linguistic competence and linguistic performance. A quick recap is needed here. In the standard PoS argument for language acquisition, it emphasises a question, which is that there are too many grammars that are compatible with the primary linguistic data and that some might be even more intuitive for learners to adopt compared to the correct one (Laurence and Margolis 2001). It seems impossible for a learner to be able to grasp the correct set of rules by domain-general learning and inferencing. Because in the case of language, it is possible that the complexity and unintuitivity of the correct set of grammar will be disguised among the indefinite set of rules. However, children do not face such issues when acquiring language.

The second reason is the learning window for first language acquisition, which is not more robust than that in other psychological learning mechanisms. The claim is that in order for a child to master the rudimentary aspects of most aspects of language, the golden period is from birth to the third year of life: before first year is prelinguistic development; after that, the development of single-word utterances takes place during year 1 to year 1.5, then the first word combination from year 1.5 to 2, and simple and complex sentences from year 3 (Ingram 1989, 2). This is a reason why it is nearly impossible to master a second language after year 6, and no other knowledge acquisition shares the same learning window as linguistic learning. I might be able to learn how to make the English breakfast after being in the UK for 7 years, but I will not be able to master English just as native speakers. The fact that the complex linguistic principles can be acquired accurately by children indicates an innate competence for acquiring linguistic knowledge. Moreover, this complex linguistic knowledge acquisition has a specific learning window which other knowledge learning does not require, which further makes it robust that this innate competence of linguistic knowledge must be domain-specific.

The innate domain-specificity of linguistic acquisition has profound implications in the practical strategy for language learning: if you want to master a language, you should start learning as early as possible, and be exposed in that linguistic environment as early as possible. For example, I have struggled with the difference between definite articles and indefinite articles, which grammatical knowledge is very intuitive to native English speakers.

However, moral knowledge do not share this similarity with linguistic knowledge, because it is unclear if moral rules and principles are that complex, and the learning strategy of moral rules doesn’t seem to require a specific method compared to learning other knowledge. Hence, the practical implication for moral learning will be fundamentally different from language learning.

Firstly, rudimentary moral rules and principles are not as complex as linguistic knowledge. What moral Chomskyans are trying to argue is that outside information and environment are not sufficient to explain the moral capacity the child processes (Dwyer 2006). But it is not clear if that is the case in moral learning. Rather the set of fundamental moral principles can be more intuitive and less complex compared to other sets of principles and norms in the community. This does not mean that individual moral principles are simple and less complex, but it is plausible that the moral knowledge that children encounter is basic and simple (Sripada 2008). There are debates on what are the universal moral norms, here I will just adopt moral foundation theory by Haidt and Graham (Haidt and Graham 2007): there are five innate and universal moral norms, which are care/harm, fairness/cheating, loyalty, authority and purity. Compared to the entrance requirements for linguistic knowledge, they are simple and intuitive. These moral norms can be acquired without learning any complex grammars to actions. By associating these moral norms with actions or being taught explicitly about the moral concepts, children can acquire and internalise these moral norms, which suffice to master moral cognition in a rudimentary sense.

The second difference lies in the learning strategy for moral learning. First language acquisition has to take place at a certain stage of life, which differs from other knowledge learning. However, this is not the case in the logic for carrying out moral learning: moral learning can take place in a longer period of life during social learning. If we acquire moral norms the way like how we acquire first language, then the moral norms we have acquired since we are children will forever stay with us, and will become the primary moral compass for our moral lives. However, there is no evidence suggesting that moral learning has to take place at a certain stage of life. This further suggest that moral learning does not require an innate domain-specific moral faculty.

Therefore, the implication for how to master moral knowledge is that we do not need to feed children a significant amount of moral data at a certain stage of their lives. In addition to providing children with the moral learning environment, we can also directly teach children the moral rules that we want them to learn. We also do not need to worry that after a certain stage of life, children will not be able to master moral values: the learning process of moral values can take place when children are learning other non-moral values.

* 1. The Problem from Mental Representation

The second difference is from the representation level. At the representational level, we ask the question of the representation of input and output, and what is the underlying algorithm that facilitates the transformation between inputs and outputs.

The difference between linguistic representation and moral representation manifests in two ways: the first is the efficiency and accuracy of inputs; the second is the algorithm(s) and processes that make the input-output transformation possible.

Let us focus on linguistic representation first. The linguistic representation of inputs, and transformation between inputs and outputs further confirms the domain-specificity of linguistic faculty from the algorithm level. Linguistic representation of inputs demonstrates fast, efficient, and reliable features that an innate domain-specific mechanism possesses. Just as the poverty-of-stimulus argument has illustrated, in a learning environment, an indefinite number of other sets of principles are consistent with the regularities found in the primary linguistic data (Laurence and Margolis 2001, 222). However, children can pick out the right grammar. This has shown that not only do we have an innate grammar for representing inputs, but also that we are very sensitive to the linguistic data in the environment. This explains why members of the same community all acquire the same language, and are able to communicate with each other.

Now let us turn to the transformation from inputs to outputs. The linguistic transformation from inputs to outputs demonstrates that this innate grammar is modular, because it is dissociable in the representational level, and it even demonstrates informational encapsulation. For example, once you have mastered the rudimentary aspects of your first language, you will probably be confident about your linguistic knowledge forever. Even if you learn a second language, the linguistic grammar of your second language will not override that of your first language competence and performance.

However, moral representation is a different story. Firstly, as regards representation of inputs, we do not seem to have a universal moral grammar that functions fast, efficiently and reliably for us to detect moral cues from our environments. If the linguistic analogy worked in the moral case, we could translate the narrative as follow: the mind contains an innate moral faculty that makes human moral representation possible. The moral faculty consists of a set of rules, principles and constraints that helps children detect moral cues from the environment. In this way, Universal Moral Grammar helps children form their mature moral competence which is moral idiolect. If this theory works, then members in the same moral community should adopt same moral norm systems. Because according to Chomskyan moral nativists, the information inputs are processed by universal moral grammar, and further facilitates individuals’ own moral systems. Since universal moral grammar is innate, each individual’s moral information processes should be the same (in other words, how universal moral grammar represents morally relevant elements should follow the same code). Meanwhile, within the same moral community, the moral cues in the environment should be the same. Since the information and the universal moral grammar are the same, we should speculate that in an isolated and independent community, individuals should only adopt one single moral system. But this is not always the case. Throughout human history, moral disagreements within one community take place. For example, civil wars happen in nearly every single human culture and country (S. Neumann 1949), and most of them are due to moral disagreements within the same society. In this sense Chomskyan moral nativists ignore that human morality demonstrates diversities *within* human groups. If there is universal moral grammar, this should not happen, as universal moral grammar constrains what can be moralised and how to moralise. Hence, there is no universal moral grammar at the representational level that makes us sensitive to the ‘correct’ kind of moral cues, and accurately detect them from environments.

In addition, there is no evidence suggesting that the transformation from moral inputs to moral outputs demonstrates dissociability and information encapsulation. There is no evidence suggesting that the first moral system you adopted will be the dominant moral values that stay with you for the rest of your life. This would sound very depressing for human society. Instead, one’s core moral system is always responsive to outside stimulus. At the same time, our moral system is subject to influences of other psychological mechanisms or other stimuli even after we developed a moral idiolect. For example, a person’s moral system can be updated when they encountered new evidence or living in a new country. I might not be able to acquire the same level of competence at English as a native speaker, nor be able to represent the grammatical rule of indefinite articles as a native speaker, but I will be able to adopt some moral values that I have learned here, or even more, I might completely update my moral system if I think the new moral value system is better.

Therefore, at the representational level, moral cognition also does not demonstrate the features that are normally manifested in an innate domain-specific mechanism like the linguistic system. Specifically, the moral system does not represent input in a fast, efficient and reliable way, and the internalised moral system is not dissociable with other psychological systems or stimuli.

* 1. Problem from Physical Realisation

The last challenge is the problem from biological evidence. This challenge is from the implementation level. At the implementation level, we ask the question of how the cognitive system is realised physically.

The cognitive and neural basis of human language has been studied extensively. Research has been conducted to investigate the emergence of language via evolution: how is our language system (in the neural sense) evolved from proto-vowels in non-human primates[[8]](#footnote-9)? So far, studies on vocalisation and speech have already reliably revealed the evolutionary trait of human language capacity and pinpointed our language capacity at a physical level (for review, see Boë et al. 2023). These empirical studies pinpoint an area for language information-processing, and also situate it in the picture of evolution by contrasting the human brain with other animals’.

For example, Broca’s area, which is located in the frontal lobe of the dominant hemisphere (normally it’s the right hemisphere), has been linked to speech production (Broca 1861; Jackendoff 1999; Del Maschio et al. 2022). Hagoort further (2005) argues that Broca’s area plays an essential role in unification of domain of language. This line of empirical research further strengthens the nativists’ claim about the nature of language: not only that the linguistic knowledge is a natural kind, but also that we can pinpoint the neural areas that realise the cognitive function of language[[9]](#footnote-10).

Furthermore, recent comparative studies on the sulcal patterns of ventrolateral prefrontal cortex across different primate species provided further explanation of the uniqueness of human language. For example, Amiez and collaborators (2023) have argued for a unique neural feature that enables human capacity for speech. This feature is called the Prefrontal extent of the Frontal Operculum (PFOp) region, which is located in the ventrolateral prefrontal cortex, adjacent and ventromedial to Broca’s area. By comparing human brains with that of chimpanzees, baboons and macaques, they find that only humans have a fully opercularised PFOp. This means that it creates a buried gyrus between Broca’s area and the anterior insula, which will enhance the capacity of speech processing. In comparison, while chimpanzees do have the precursor of PFOp, monkeys do not have any form of PFOp. They further argue that PFOp is a result of speech evolution: when faced with increasing selection for cognitive and motor functions related to language, humans evolved this PFOp to speed up and facilitate the linguistic representational process, which is crucial for realising the linguistic competence of humans. An oversimplified example might be helpful with understanding the jargon: a manifestation of our linguistic competence and universal grammar is that we can detect the linguistic elements from numerous sounds and noises in the environment. Emergence of PFOp functions as efficiently selecting the language-relevant inputs by filtering non-linguistic elements and bridging the Broca’s area with representation of linguistic inputs. However, non-human primates are not endowed by this linguistic mechanism.

Hence, this demonstrates why humans can be sensitive to linguistic cues: we are able to differentiate linguistic cues from non-linguistic sounds because we have a ‘designed pathway’ that can speed up the process of linguistic representation from input to output. This pathway not only explains our linguistic capacity from the implementation level, but also situates this level of explanation in the evolutionary picture.

While neuroscience has provided exciting news to human linguistic competence, it looks somehow depressing for moral competence. Experiments done by Parkinson and collaborators show that distinct neural systems underly different moral judgements (Parkinson et al. 2011). Parkinson and her collaborators focus on three kinds of moral wrongness judgements: disgust-based moral judgements, harm-based moral judgements and honesty-based moral judgements. They found out that distinct brain regions, which are responsible for different cognition, are recruited. Brain regions associated with mentalising, affective processing, and action understanding are activated respectively when participants are making moral judgements about disgust, harm and dishonest accordingly. This has shown that at the neurological level, we cannot pinpoint a unified moral faculty. One might refute the conclusion by referring to the plasticity of our brain (Demaree-Cotton and Kahane 2018): even the Fodorian researchers allow multiple realisability. Modularity of mind does not exclude the possibility that for the same cognitive function, it can be realised by different physical devices (Carruthers 2006). This could be a valid concern, but the problem remains for moral Chomskyans: even if we accept that the process of naturalising properties does not require a story from neuroscience, moral Chomskyans still need to provide another solution to frame the domain-specificity of moral faculty.

* + - 1. Conclusion

Therefore, based on my criticism at section 3, at all three levels of understanding a system, the linguistic system and the moral system do not share similarities. It should suffice to conclude that the linguistic analogy argument fails, and there is no such thing as universal moral grammar.

In this chapter, I have argued against Chomskyan moral nativists by refuting the linguistic analogy. It is worth re-emphasising that this chapter does not tend to reject Chomsky’s linguistic theory, and I remain neutral to the view that the mind is modular in character, and the mind contains diverse cognitive structures defined by its specific properties and principles. It is possible that knowledge of language is acquired, processed and executed by this innate domain-specific module. My view is that moral cognition is not achieved by this modular mind, because our moral system does not exhibit the key features that a modular cognitive system possess. The key features for a modular cognitive structure include a set of rules and principles devoted to that specific psychological function. This feature entails exclusivity and encapsulation, and morality does not manifest this feature.

One may argue that my criticisms of Chomskyan Moral Nativism are not charitable enough. Defenders of Chomaskian moral nativism would contend that the analogy between language and morality is not in a very strict sense: it is not the case that the moral system is *exactly* the same as our linguistic system. The analogy is a loose metaphor to help us understand the general structure of moral cognition, but it does not entail that the operation is the same. However if the analogy between language and morality is only superficial, then there is no theoretical implication from the analogy.

## Chapter 5 Cooperation and Morality: Cooperation is Not the Whole Story of Moral Cognition

In chapter 4, I have looked at an evolutionary approach to moral cognition – Chomskyan moral nativism, arguments for which rest upon Chomskyan linguistic knowledge. I have argued that moral knowledge does not demonstrate the same features as linguistic knowledge, hence linguistic analogy for moral nativism fails. If you commit to the claim that there is an innate moral faculty result from evolutionary adaption, another popular strategy is to assert that morality stems from cooperation.

In this chapter, I will look at this alternative account for innate domain-specific moral cognition: cooperation-based mutualistic account of morality. According to theories of cooperation-based morality, morality stems from cooperative concerns, and cooperative concerns are part of human psychological traits as a result of evolutionary adaptation (Sperber and Baumard 2012; Baumard, André, and Sperber 2013; Cosmides, Guzmán, and Tooby 2018; Tomasello 2020). Therefore, moral cognition emerged because we need to facilitate cooperation with others, and we need to evaluate people’s moral characters in order to assess whether they are suitable for cooperation. Furthermore, some argue that moral cognition is realised by a unified and domain-specific mechanism, because morality is essentially about fairness representation in cooperation(Baumard, André, and Sperber 2013; Fitouchi, André, and Baumard 2022).

In this chapter, I will tackle the mutualistic approach. I will argue that they fail to unify moral cognition in cooperation concerns, even if they broaden the concept of ‘cooperation’ to fit moral cognition in, the problem still can’t be solved, as moral representation is not just about fairness representation.

Section 1 will be an introduction of cooperation-based accounts of morality and mutualistic model of morality. I will unpack two of their arguments: the argument from ‘morality is fairness’ and the argument for puritanical norm. The first argument I will consider is their central argument that moral cognition can be grounded in cooperation especially fairness representation, which is sufficient to infer that moral cognition is a unitary faculty that results from cooperative concerns (Baumard, André, and Sperber 2013; Fitouchi, André, and Baumard 2022). After considering their central argument, I will demonstrate one specific argument on puritanical morality they provide for the unitary account of moral cognition: Fitouchi and collaborators have realised the potential challenges of the mutualistic model, one of which is the existence of puritanical norms. I will consider two of their approaches: in their first approach, they carry on taking evolutionary psychology’s approach, and argue that puritanical norms emerged and exist stably in human society because we need to assess other people’s *self-control,* which is a crucial virtue for cooperation. According to Fitouchi et al. (2022), people have the folk psychological belief that people who are good at self-control is trustworthy for cooperation. Hence, puritanical norms are still fairness representation concerning cooperation. In their second approach, they explain the variety puritanical norms across cultures by referring to *cultural evolution*, which is the view that cultural and social learning shapes our cognition similar to how biological evolution do. The change of cultural and social environments will shape our moral cognition, including puritanical morality.

Section 2 will be my criticism of mutualistic approach, and I will argue that mutualistic model is wrong about moral cognition. I will focus on Fitouchi and collaborators’ argument for puritanical norms. Once I prove that puritanical norms are irrelevant to cooperation, this will serve further to refute the unitary account of cooperation-based moral cognition. I will argue that puritanical norms are not for the assessment of self-control and trustworthiness of a person, and self-control is not an important factor for trustworthiness of a person. At the same time, self-indulgent behaviours are often perceived as promoting cooperation within group. There are two challenges for claiming puritanical norms as cooperation concerns: the first is on the relationship between self-control and trustworthiness for cooperation: self-control is not always conceived as a good virtue a good partner has. Studies have shown that people also have negative impression of people with high self-control. Moreover, many practices of puritanical norms (practicing self-control) are encouraged to take place alone without being noticed by others, hence the link between self-control and trustworthiness for cooperation breaks. The second criticism will be on the relationship between self-indulgent behaviours/puritanical norms and trustworthiness for cooperation. If Fitouchi et al. is right about puritanical norms, self-indulgent behaviours such as drinking and dancing should be forbidden in societies which the resources are scares. But this is not the case. Therefore, puritanical norms must emerge from other concerns. The existence of puritanical norms should be enough to indicate that moral cognition is not unified by cooperative concern.

After rejecting mutualistic model of morality, section 3 will further assess the evolutionary psychology approach as well as cultural evolution. Therefore, this section can be considered as a wrap up for the implications from both chapter 4 and chapter 5. I will argue that evolutionary psychology has indeed provided a useful tool to understand the architecture of moral cognition. However, the theory itself cannot justify the domain-specificity of moral cognition. Meanwhile, I will also discuss the role of cultural evolution in moral cognition, since many moral theories, including the mutualistic model implies the influence of cultural evolution(Boyd and Richerson 1985; Boyd 2019) in moral cognition. I will argue that the cultural evolution framework does not help the moral nativist: arguably cultural evolution is not helpful with understanding any cognitive capacity. Because cultural evolution itself faces two serious problems: firstly, the similarity between culture and biology is just as shallow as the similarity between linguistics and morality. It demonstrates in several ways: their goals (computational level), their inheritances (representational/algorithmic level), and their physical realisations (implementational level). Secondly, compared to evolutionary psychology, it does not explain more human behaviours and psychologies, which make this theory redundant. Therefore, cultural evolution will not be a helpful tool for moral nativists.

1. Preliminary Work: Morality as a Discipline for Cooperative Concern

In general, there are two theoretical frameworks that inform cooperation-based moral theories: one is evolutionary psychology, and the other is economic game theory. Evolutionary psychology is the grounding of cooperation-based theory. Cognitive evolutionary psychology is the proposal that human cognitions are results of adaptation to nature in order to survive and reproduce (Kelly and Setman 2021; Tooby and Cosmides 2005). This branch of theories shares a claim: namely that the essential cognitive capacities that human beings have are the result of successful adaptation of the environment and challenges human ancestors have faced. Moral cognition is an example of the evolutionary achievements that human beings have had. This makes humans distinct from other animals: While other animals such as chimpanzees like achieving goals alone, humans at the early age have shown that they like pursue goals by collaborating with others (Warneken, Gräfenhain, and Tomasello 2012).

Having the evolutionary ground, theorists of cooperation-based morality build up their theory based on assumption that human interactions are for cooperation ultimately, and they use ‘non-zero-sum' game theory to frame the structure of their theories. ‘Non-zero-sum' game theory is a model that describes how people (players) make decisions in the competitive or cooperative scenarios where choices of both parties affect each other, and rewards and costs experienced by all parties do not balance. In other words, one player’s gain is not necessarily another player’s loss, and the relationship between two players are not necessarily competitive like sports game. Cooperation-based moralists view human interactions as a whole, parallel to what happen in ‘non-zero-sum' game. In other words, they perceive human interactions as cooperation, and people build and maintain social relationship because they want to facilitate cooperation. Following this line of thought, social and moral norms stem from securing cooperative functions.

Building from evolutionary psychology as well as game theory, several cooperation-based moral theories have been proposed (Baumard, André, and Sperber 2013; Curry 2016; Earp et al. 2021). There are two kinds of cooperation-based moral theories based on their different presumption of moral faculty: the first one is called the problem-centred approach to morality. This approach does not commit to the domain-specificity of moral cognition and argues that there are multiple goals for cooperation(Curry 2016). Because cooperation entails distinct functions and goals, the psychological mechanisms vary. This naturally leads to the rejection of domain-specific moral faculty; another kind of theory is known as the mutualistic model to morality. According to mutualistic model to morality, all moral cognitions are about fairness representation, which is sensitive to mutual benefits between cooperative partners(Baumard, André, and Sperber 2013; Fitouchi, André, and Baumard 2022). In this chapter, I will specifically focus on the mutualistic model rather than problem-centred account. Because I do not deny that cooperation is essential to moral cognition. What I am contesting is the claim that by grounding morality in cooperative concerns, we can further infer that moral faculty is unified.

* 1. Central Claims: Moral Representation is about Fairness Representation

The main claim of the mutualistic model is simple: the need to cooperate with others is why we have an innate domain-specific faculties. But there are more to unpack. According to Baumard and collaborators, morality is about fairness representation (Baumard, André, and Sperber 2013, 77):

‘…the argument developed throughout this article that a specific and non-instrumental preference for fairness evolved as a distinct “moral sense.” If you favor a more extensive definition of morality, call this a “fairness sense.”’

As can be seen, to say that morality is the result of cooperative needs is not enough. Because a pluralist cooperation-based theorist also commits to this claim. Hence, as a unitary moral theory, mutualistic models asserts that morality emerges from a concern for fairness.

But why is cooperation important in the first place for human society? This is where evolutionary psychology comes in. According to evolutionary psychology, humans are social animals facing the challenges of adaptation and fitness to the environments. The evolutionary drive equipped us with an inherited psychological preference for cooperation with others (Tooby and Cosmides 2005). Since cooperation is essential for humans, relevant social and cognitive systems emerged to facilitate cooperation among people, one of which is a sense of fairness. The concern for fairness sparks three core moral forces when we are interacting with others in our society: to care about reputation for cooperativeness(Sperber and Baumard 2012), to punish people who hinder cooperation (Baumard, André, and Sperber 2013), and to assess others for our partner choice (Tomasello 2020).Therefore, the function of moral cognition is fairness representation. This process can be seen as a process of cultural evolution: the exact function and representation of fairness is formulated during social learning and social interaction.

* 1. Solving the Problem of Puritanical Norms

The mutualistic model has faced several problems. But in this chapter, I will just focus on one problem: the mutualistic model cannot explain the existence of puritanical norms.

An interesting phenomenon in human morality is that we have a set of moral norms that condemn apparently victimless and harmless behaviours such as drinking, feasting, dancing and having sex too much. In contrast, we also emphasise the virtue of self-discipline and pious lifestyle. This pair of norms is known as puritanical norms. The existence of these norms have raised a big challenge for mutualistic account. Because while puritanical norms are moral norms, they do not seem to be relevant to cooperation functions. It seems to be the case that cross-culturally, puritanical norms exist, and societies condemn people who engage self-indulgent but harmless behaviours such as taking (harmless) drugs, drinking, feasting, wearing immodest clothes, and even dancing. At the same time, societies also promote pious behaviours. The persistence of puritanical norms across human history and cultural groups suggest that it is really a kind of moral norm(Fitouchi, André, and Baumard 2022). However puritanical norms are about regulating one’s own behaviours, which is irrelevant to cooperative concern. For example, whether I am going out for a party should not affect my reputation of whether I am a good partner for cooperation.

Hence, if you are a mutualistic theorist who also argue for a unitary account for moral cognition, you need to provide an explanation on how to include puritanical norms in our moral domain, and how to explain the changes of puritanical norms over time and across cultures. Fitouchi and collaborators take the challenge: they argue that puritanical norms are for the assessment for one’s self-control, which capacity is considered crucial in cooperation.

At the same time, they argued that people have two folk psychological beliefs about behaviours, self-control and cooperation: People hold the folk psychological belief that engaging in self-indulgent behaviours, such as drinking, dancing, and feasting, will amplify the motivational force of short-term cravings, leading to a lack of self-control; at the same time, people also hold the folk psychological belief that a lack of self-control hinders cooperation. They argue that these two folk psychological beliefs are results of evolution: we need a detecting system to help us quickly assessing a good cooperative partner, especially when the resource is limited in the society. By linking puritanical norms and cooperation with self-control, their argument is formulated as follows(Fitouchi, André, and Baumard 2022, 14):

1. People hold the folk psychological belief that engaging in self-indulgent behaviours, such as drinking, dancing, and feasting, amplifies the motivational force of short-term cravings, leading to a lack of self-control.
2. People also hold the folk psychological belief that a lack of self-control hinders cooperation.
3. As a result, puritanical norms, which prohibit such self-indulgent, yet apparently harmless, behaviours, arise out of a need to preserve self-control which is essential for cooperation.
4. Moral cognition is indeed grounded in a need for cooperation.

The above is their general arguments. As can be seen, this argument has been influenced by the theory of evolutionary psychology, especially on how human’s early tribalist society can shape human cognition. The key feature of early human tribalist society is ‘poor’: the resource for reproduction and survival is highly limited. Hence a good cooperation and a good cooperative partner are matters of life and death. These two folk psychological beliefs, and puritanical norms emerged under this background: we need to find out partners that trustworthy for cooperation.

At the same time, they also acknowledge the fall of puritanical norms in some societies. Fitouchi and collaborators took cultural evolution approach to explain why puritanical norms are universal but has declining particularly in the western, educated, industrialised, rich and democratic societies(Fitouchi, André, and Baumard 2022, 34). They argue that social change place a crucial role in the change of puritanical norms. The original puritanical norms emerged because people need to identify trustworthy partners for cooperation in societies where all resources are limited. Based on study from Nettle and Saxe (2020), Fitouchi et al. further argue that this poor society lack trust between people. In a society where the resource is limited, it is crucial to make sure that the cooperation goes well. In order to facilitate cooperation and promote better utility, it is crucial to know that the other party is trustworthy. However, in a poor society, due to the low level of trustworthiness in society, people need another reference to assess the potential co-operator. According to Fitouchi et al., self-control is a good measurement: if a person is good at self-control, it suggests that this person is trustworthy for cooperation. And the best way to test one’s self-control is by observing whether they are good at preserving from self-indulgent pleasure. Hence puritanical norms emerge to assess one’s self-control and potential for cooperation. This also explain why puritanical norms decline in WEIRD societies: those societies have higher trust among people, and it is not necessary for people to refer to others’ self-control to evaluate others’ trustworthiness in cooperation.

1. Evaluation

In previous section, I have considered the argument on puritanical norms provided by mutualistic model. According to mutualistic model, puritanical norms also stem from cooperative concerns. In this section, I will argue that they are wrong about the relationship between self-indulgent behaviours (puritanical norms), self-control and trustworthiness for cooperation.

Firstly, Fitouchi et al. are wrong about the relationship between self-control and trustworthiness for cooperation. which is complicated than Fitouchi et al. have suggested. Studies on self-control reflects a complex impression people have for self-control. Admittedly, studies that have been cited by Fitouchi and collaborators in their paper might probably show that people who are good at self-control seem more trustworthy. However, the evidence they have cited is about lay people’s impression of religious people who practice pious lifestyle(Moon, Krems, and Cohen 2018). However, it is possible that people have formed two impression of religious people, one is being trustworthy, and another is being good at self-control. However, it is possible that these two impressions are not causally related. Another explanation is that the trustworthiness in Moon et al.’s experiment does not function the same as what Fitouchi et al. mean by trustworthiness. Yes, we might think that a religious person who practices pious lifestyle is can be trusted. But it is not the same idea as being trustworthy as a cooperative partner. Even if we accept the data from Moon et al., other studies also show a completely different story, and this might further show that we do not have a coherent belief pattern for self-control and trustworthiness for cooperation. High self-control can potentially harm one's social image and reputation. Studies have shown that individuals with high self-control are often perceived as robotic and unsympathetic. People who speak and act impulsively, on the other hand, are generally seen as more honest in comparison to those with high self-control(Lapka et al. 2022). Research conducted by Röseler and colleagues also indicates that individuals with high self-control are often perceived as unreliable, largely because they are seen as lacking power and status (Röseler et al. 2021). Being perceived as robotic, unsympathetic, and unreliable can hinder cooperation, given that individuals typically prefer to collaborate with those who appear more human, sympathetic, and reliable..

Secondly, they are wrong about the relationship between self-indulgent behaviours and cooperation. It is not the case that people have the psychological belief that self-indulgent behaviours corrupt self-control and damage cooperation. Quite opposite, evidence show that collective self-indulgent behaviours can facilitate trustworthiness among others, because by enjoying collective self-indulgent behaviours together, people build up emotional bonds and community identities. To better justify my claim, I will refer to evidence from archaeological studies. If their theory is correct, puritanical morality should exist at the early stage of human societies. Because when a community just emerges while resources are limited, according to Fitouchi et al., people would not build up trust to each other. Therefore, people should alternatively apply puritanical norms to assess others’ self-control. However, archaeological evidence points to a different story: people do not perceive self-indulgent behaviours as hindering cooperation, rather, people think that activities such as drinking, dancing and feasting facilitate cooperation.

For example, During the Neolithic age (ca. 8000-3000 B.C) and Bronze age (ca. 3000-1200 B.C), China underwent a rapid population growth. As populations grew there was a greater need for cooperation as there was a scarcity of resources. This growth coincided with the domestication of plants and animals, and communal farming, tasks which required cooperation within family groups and communities. At the same time there was the emergence of larger public buildings (Crawford, Inbar, and Maloney 2014; Bestel et al. 2018; Liu 2021). During this period, with population growth there was a shift in the organisation of settlements. Public houses, located in village centres, increased in size and smaller family dwellings were built to surround these public centres. As part of these village centres, large communal hearths were constructed for communal cooking. Large, decorated, amphorae for the sharing of alcohol were also located in these centres(Liu 2021). During this period, as populations grew there was a substantial investment of time and resources for creating the means for social interactions centred around drinking and feasting. These studies show that alcohol consumption facilitated cooperation in three ways: (1) it helped construct community identity and community hierarchical structures; (2) it helped maintain mutual aid networks between households; (3) it promoted intracommunity cooperation. If we follow Fitouchi et al.’s account, puritanical norms should have existed during this period. People should condemn members in the community who like drinking and feasting together. However, evidence show that people not only didn’t condemn drinking and feasting, instead, people have spent time and efforts in participating and facilitating drinking and feasting events.

As can be seen from my arguments against Fitouchi and collaborators, puritanical norms are not for assessing others’ self-control, therefore puritanical norms are not for facilitating cooperation. This is because self-control is not always considered as a candidate for good cooperative partners.

1. Problem of Cultural Evolution and Evolutionary Psychology

In the last section, I have argued that mutualistic model fails to ground all moral cognitions in cooperation, In this section, I will further examine the theoretical framework they rest on: evolutionary psychology and cultural evolution.

I have argued in chapter 4, that, when moral Chomskyans try to use evolutionary psychology to justify moral nativism, evolutionary psychology will disappoint them. The argumentative strategy they took was to make an analogy between language and morality. Initially, this is a very good strategy, because if morality is really structurally like language, we would find a very reliable reference for studying morality as a computation system. Because linguistics has been extensively studied in all three levels (this is not to say that we have solved the problem of all the linguistic problem). Even though there are ongoing debates on whether Chomsky is right about modularity of language(Fedorenko and Varley 2016), the study of linguistics has been progressing fast.

In contrast, mutualistic theorists build up their account by referring to another kind of theoretical framework: cultural evolution. There is an ongoing debate on contents of cultural evolution: what are the cultural inheritances, and what have evolved during cultural evolution. Although labelled ‘evolution’, cultural evolution goes parallel to biological evolution(Lewens and Buskell 2023). In the following, I will cast doubt on the culture/biology analogy for the existence of cultural evolution and its explanatory role in moral cognition: I will argue that there is no evidence suggesting that cultural ‘evolves’. If there is no such thing as cultural evolution, it will inevitably threaten the soundness of the theories that rest upon cultural evolution.

The first and the most important difference lies in the computation level. from the computational level, biological evolution’s goal is clear: reproduction and survival. In order to reproduce and survive, an organism should be good at demonstrating adeptness and fitness to the environment. Building up on biological evolution, evolutionary psychology can infer that our behaviours and way of thinking are shaped by the process of evolution. However, this is not the case for cultural evolution. Even though we might have a general idea about what cultural learning mean (normally mean the process of social learning and learning new knowledge from other). However, this is not the same for cultural evolution. Firstly, what properties evolve in the case of culture? does culture really evolve, or it just simply change? What is the goal of cultural evolution There are valid reasons for asking these questions about cultural evolution. Because there is no clear idea of what elements have fundamentally changed during cultural evolution. Let us use puritanical norms as example again, puritanical norms change of course, but it only has limited status in given cultural environments: either they exist, or they are not followed anymore, or they become more conservative or liberal. Admittedly, a puritanical norm might become another kind, for example, the norm that you should take a shower can be a conventional one and can transit to a moral one, depending on the individuals. But the change does not reflect any inherent change in content. Meanwhile, it is unclear what is the goal of cultural evolution. Of course, one can just borrow evolution’s theory to say that cultural evolution is for adapting to the social environment. However, it is unclear whether it has the urge to go through evolution.

Secondly, from algorithmic level to compare, evolution has already been able to pinpoint the decipher for information processing and the genetic codes that carry out the (genetic algorithm for example). At algorithmic level, we ask the question of what is the representation of inputs, representation of outputs and how is the transformation between representation of inputs and outputs possible. Thirdly, physical realisation level, while evolutionary psychology can locate to gene and DNA sequence to physically realise the information, this is not the case for cultural evolution.

Therefore, from Marr’s three-level to compare both biological evolution and cultural evolution, we can see that just like linguistic analogy for universal moral grammar, the biology/culture analogy is superficial: from the computational level, it does not have the clear goal and functions as biological evolutions do. Because in the social domain, we do not have the principles and parameters to measure the adaptness and fitness of a culture, and we probably do not need one. At algorithmic level and implementation level, while biology is able to provide evidence for sketching out the evolutionary trait of natural selection, cultural evolution fails to have the same promise at algorithmic level and implementational level.

Moreover, there are potentially dangerous consequences if we commit to the claim that cultural evolution is for adapting to the social environments. If that is the case, we would have a measurable parameter to evaluate whether a culture is better and stronger than the other one from the aspect of cultural evolution, meaning that we can compare two cultures from an objective point of view: a culture is moral successful if it is more widely accepted by most people. Firstly, I do not think that this is objectively true. Admittedly, there are some cultures that appear more dominant and widely accepted by people, but it does not suffice to suggest that it is a better form of culture compared to the alternative one. For example, during the colonialism, a politically and militarily powerful country directly controls less powerful countries. Normally, the culture from the powerful country will also compete with local cultures, and most of the time it become the dominant culture in the local area. If the parameter for cultural evolution is right, then a potential consequence would be justifying the dominance of colonial culture in local culture from the aspect of cultural evolution.

The second issue for the cultural evolution theory is that it does not increase a better understanding of questions we concerns in social domain(Sober 1992), this also applies to understanding questions we concern in moral domain: quite opposite, maybe theories of evolution only add another layer or another jargon between the question and answer. For example, in moral psychology, we ask the question of why humans adopt some moral norms not others. For example, we ask the question of why human adopt the puritanical norm that drinking is wrong, and why this norm has been widely accepted across cultures, but at the same time it has been abandoned by some cultures nowadays. If you approach to this question from theories of cultural evolution you might come up with an explanation with social learning from others in your environment in conjunction with a built-in social learning system. Fitouchi and collaborators took this approach to explain why puritanical norms are universal but has declining particularly in the western, educated, industrialised, rich and democratic societies(Fitouchi, André, and Baumard 2022, 34): while puritanical norms emerged from early human society when resource are scares, those societies have higher trust among people, and it is not necessary for people to refer to others’ self-control to evaluate others’ trustworthiness in cooperation. But the problem is: why do they need a theory of cultural evolution here in your explanation? It doesn’t seem necessary to apply cultural evolution to understand how puritanical norms change due to the change of social structure and social resource. Simply by referring to evolutionary psychology, it is already enough to explain the emergence of that cognition. And in terms of explaining how social contexts shape puritanical norms, the causal relationship does not need to go cultural evolution.

1. Conclusion

In this chapter, I have demonstrated another account—cooperation-based mutualistic account, which ground moral cognition in cooperative concerns. I have argued that cooperation is not the whole story, and just like moral Chomskyanism, mutualistic approach to morality fails to provide a convincing argument for a unified moral domain.

Specifically, mutualistic approach theory argues that moral cognition emerged to track fairness in cooperation, assess one’s trustworthiness for cooperation, and maintain reputation as good co-operator. This account also faced several challenges. And in this chapter, I have demonstrated one of the biggest challenges they face: to unify puritanical norms in cooperative concerns(Fitouchi, André, and Baumard 2022). Fitouchi and his collaborators argue that puritanical norms stem from assessing one’s self-control, which is crucial for cooperation. In section 2, I have argued that their argument fails to connect puritanical norms to cooperation with self-control. Following, in section 3, I have discussed the implication of evolutionary psychology as well as cultural evolution in moral cognition, since mutualistic approach implies that evolution and social learning work together to promote moral cognition. However, I argue that neither evolutionary psychology nor cultural evolution can further prove that moral cognition is innate and domain specific.

# Part 3: General Learning Mechanism, Machine and Morality

In part 1, I have discussed two accounts of moral sentimentalism (chapter 1 and chapter 2), I have argued that we cannot ground the psychological mechanisms of moral cognition in affective mechanisms. In chapter 3, I have discussed the role of emotions in moral cognition. I have argued that while emotions do influence moral cognitions, empirical studies have shown that emotion’s influence on moral cognition does not suggest that emotion is always the compass for moral judgements, and it is questionable whether evolution can explain the emergence of moral emotion as well. This has two implications: the first is that emotion is not the whole story of moral cognition, and the second is that the emotion/reasoning dichotomy is not a good way to understand moral cognition, because it is normally the case that the psychological mechanism(s) that facilitate(s) moral cognition, no matter whether they are innate or not, might include both affective and reasoning components.

In part 2, I have examined two theories that take another approach: this approach more or less remains neutral on the traditional sentimentalism/rationalism debates, instead, it assumes that a distinct psychological faculty is responsible for moral norms acquisition. This capacity is a result of evolution. Chomskyan moral nativists make an analogy between language and morality, and argue that moral congition requires a moral grammar for moral norms acquisition. I have argued that the linguistic analogy approach fails to pinpoint the psychology of moral cognition, because the way we learn and use moral norms is fundamentally different from that in language. mutualistic moral theory draws attention to the needs for cooperation in human society, and argues that moral cognition emerged to guarantee better cooperation in community. I have argued that mutualistic moral theory also fails to prove the existence of a domain-specific moral faculty. There are two main problems for this big leap: the first is that we have moral norms that are not relevant to cooperative concerns, and the second is that even if cooperation was the major concern when moral norms firstly existed, it is not clear whether this is too distal and it does not tell us how our moral cognition works when we have already picked out these moral norms. Following, in chapter 6, I have also discussed the relationship between evolutionary psychology and morality. I have argued that evolutionary psychology is indeed appealing. Just like other human psychology and behaviours, human moral psychology and moral behaviours are shaped by evolutionary process. However, it does not further suggest that a domain-specific moral faculty is necessary.

Hence, all the discussions so far suggest that moral capacity is realised by a *general* psychological mechanism: this psychological mechanism is also responsible for other kinds of learning and cognitive processes, and the information that is represented as morally relevant is also open to other kind of cognitive representation.

## 

## Chapter 6 Problems with Moral Bayesianism

In chapter 4 and chapter 5, I have discussed Chomskyan moral nativism and mutualistic moral theory. Both models try to provide a unified and coherent psychological mechanism for moral cognition. However, I have argued that both approaches have failed to prove that a unified psychological mechanism is required for moral cognition. I have further examined whether cultural evolution can lead us to any further implications about how we make moral judgements. It seems that moral capacity is not unified psychologically, and cultural evolution, or simply referring to evolutionary psychology, cannot tell us how we make moral judgements. However, the key question remains: if there is no such thing as moral faculty, how is moral capacity implemented psychologically? What psychological components are involved? In this chapter, I will explore a model which tries to give evidence that moral norms can be acquired via general learning capacity, especially through Bayesian inferencing. This account is known as the rational rules account proposed by Nichols. Arguably the most important contribution of Nichols’s rational rules account is to bring Bayesian inferencing into the discussion of moral inferencing, and to show that moral cognition does not need a specific psychological system to be achieved.

The first part will be clarifications for key terms such as rationalism and Bayesian inferencing. Moral rationalism can indicate different theories. Hence it is important to situate the rational rules account in moral philosophy debates. Moreover, since the rational rules account grounds its theory in Bayesian inferencing, it is helpful to have a better understanding of what Bayesian inferencing is. To better walk us through Bayesian inferencing, I will demonstrate how Bayes’s theorem is derived mathematically following two examples. At the same time, I will explain why Bayes’s theorem reveals that Bayesian inferencing is essentially a rational process, and how it is different from other kinds of rational learning processes.

The second part will be the main body of this chapter: the rational rules account. I will focus on the demonstration of some key arguments from the rational rules account, especially how Bayesian inferencing enables agents to acquire moral norms and determined the scope of moral norms. I will unpack Nichols’s argumentative strategies: his criticisms of other moral theories, and Nichols’s arguments and empirical evidence for rational rules’ account and how it is different from other moral rationalisms. I will focus on one important argument that Nichols establishes: Bayesian learning excels in explaining how people acquire and identify norms in their environment. Specifically, Nichols and his collaborators have provided empirical evidence to show how Bayesian inferencing helps participants to determine the scope of a moral norm.

After demonstrating the rational rules account and Nichols’s arguments, in the third part of this chapter, I will evaluate his arguments and the empirical evidence he and his collaborators have provided in favour of moral Bayesianism. I will argue that Nichols is right that value representation cannot explain moral judgements, because value representation does not sufficiently entail moral representation, as moral representation always involves representation of abstract concepts such as harm, and moral representation is complex which is determined by the complexity of moral rules (Nichols 2021, 36). However, I will further argue that just as Nichols criticises other rational accounts because they fail to pick out the unique and complex features of moral norms, it is unclear whether Bayesian learning itself is the cognitive process for moral cognition in humans. There are three reasons to challenge the rational rules’ account: firstly, just as how Nichols criticises other moral theories because their account cannot explain the complex structures of moral norms, neither does the rational rules’ account successfully pinpoint how individuals acquire moral norms from their learning environments. Secondly, some key factors that are indicated by Bayesian inferencing, such as size principle and sensitivity to evidence, are not robust in moral cognition. Thirdly, even if we grant that Bayesian inferencing is in principle able to explain how a creature could acquire moral norms, it is still questionable whether real human beings actually use Bayesian inferencing process to acquire moral norms. The third worry is an exclusive worry for the rational rules’ account, because the requirement to be a moral Bayesian is too high: not only that a moral learner is rational and is not affected by irrelevant biased mindset, but also that a moral learner needs to be sensitive to evidence, and they are willing to update their evidence accordingly based on new evidence. Hence, compared to other accounts, the rational rules account is more demanding cognitively. I will argue that this is not the case in lay moral cognition. Hence it is possible that moral Bayesian is a good normative ethics principle, but it is not a descriptive account for moral cognition.

1. Introduction: Moral Rationalism, Bayesian learning

Before digging into Nichols’s rational rules’ account, it is important to clarify some key relevant concepts: moral rationalism and Bayesian learning, and how they are relevant to the rational rules account.

A short clarification is that the rational rules account is directly influenced by Bayesian inferencing, especially in terms of Bayesian understanding of beliefs and evidence: the rational rules account treats the formation of moral belief and moral knowledge as involving consideration of probabilities of the moral beliefs to be true given the evidence and agents’ sensitivity to evidence. Meanwhile, although sharing the same word ‘rational’, the rational rules account does not share a close connection with classic moral rationalist theories. While classic moral rationalist theories emphasises the relationship between moral reasons and action, Nichols’s rational learning account focuses on how moral rules are directly relevant to evidence an agent has. I will explain in the following paragraph.

First, moral rationalism. The term ‘moral rationalism’ picks out distinct theories due to the different understandings of moral reasons (Schroeter and Jones 2018). But in general, in philosophical discussion, moral rationalism emphasises the relation between moral principles and *justified reasons* for acting based on that moral principle (Smith 1994). This is in contrast to moral sentimentalism which generally argues that moral judgements are mainly expressing speakers’ attitudes and are not truth-apt. Moral rationalism contends that moral judgements are belief-like and can be either true or false essentially. In other words, for most moral sentimentalists, moral judgements are expressions of one’s affective attitudes, while moral rationalists argue that moral judgements indicate one’s *reasons for* moral actions, and what constitute moral judgements is not attitudes, but one’s moral beliefs or moral knowledge, and moral facts are reasons for agents’ actions (Smith 2018). Arguably, many contemporary moral rationalists are influenced by Kant’s moral theory, which argues that moral judgement is *a priory synthetic* judgement (Kant 1785). This means that moral knowledge is a priori knowledge and universally true. This metaethical commitment tends to make moral rationalists moral non-naturalists, meaning that moral rationalists normally imply that moral properties are not natural properties, because moral truths are immune to empirical refutation (Smith 2018). There are disagreements among moral rationalists on whether moral reasons belong to a kind of reasons for action (for example, whether the reason for you to judge that harm is morally wrong is the same kind of reason of you wearing a raincoat when you need to go outside while it is raining), but just as Smith has helpfully pointed out (Smith 2018), classic moral rationalists are willing to grant that moral facts are impartial and unconditional (while they might disagree on what warrant the impartiality and unconditionality of moral facts). Therefore, a domain-specific moral faculty is warranted in moral cognition according to moral rationalism[[10]](#footnote-11).

Now let’s talk about Baye’s theorem which is the theoretical framework that Nichols’s rational learning account rests upon. Bayes’ theorem describes a statistical model intending to quantify and describe the probability of an event, a proposition, or a hypothesis being true or false, based on prior knowledge of evidence that is related to the event.

Now let us consider it in detail. Generally, belief is considered as only having two value states: a belief is either true or false. If you believe that the University of Sheffield is in the UK, then either the University of Sheffield is in the UK and your belief is correct, or the University of Sheffield is not in the UK and your belief is wrong. Thus, there is no in-between state. According to Bayesian theorists, beliefs can come in degree, and a better way to understand belief is to accept that belief describes the probability/likelihood of a proposition being true or false, which directly relates to subjective confidence of an event/hypothesis. Your subjective confidence is affected by the evidence you have. So, the belief that the University of Sheffield is in the UK comes in degrees for different agents. If you are not interested in geography of the UK and know little about Academia, Bayesian might say that you assign a probability of 0.3 to the claim that the University of Sheffield is in the UK.

Let’s move onto Bayes’s Theorem. Bayes’ theorem can be mathematically structured as the following equation:

The formula itself is simple and elegant, but it might not be intuitive to most people. Hence let us understand it with a little help of two examples. The first example is for showing how we can come to Bayes’s Theorem in the first place. Let us say that I have drawn the ace of diamonds from a standard deck of 52 cards, and I want to know the probability of me drawing a king card next given that I have drawn the ace of diamonds. Let us use A means ‘drawing the ace of diamonds’, B means ‘drawing a king card’. Hence P(A) means the probability of drawing the ace of the diamonds, P(B) is the probability of drawing a king card, P(A|B) is the probability of drawing the ace of diamonds given that I have drawn a king card in the first round, P(B|A) is the probability of drawing a king card given that I have drawn the ace of diamonds in the first round, P(A∩B) is the probability of drawing the ace of diamonds and then drawing a king card, and P(B∩A) is the probability of drawing a king card and then drawing the diamond A card.

Now let us define our probabilities based on math and statistics:

1. There is only one ace of diamonds in a standard deck of 52 cards, hence in general, the probability of drawing the ace of diamonds is P(A)=.
2. There are 4 king cards in a standard deck of 52 cards, hence in general the probability of drawing a king card is P(B)=.
3. The probability of me drawing the ace of diamonds given that I have drawn a king card is . This is because if I know that the first card I drew is a king card, then there are 51 cards remaining and still only one the ace of diamonds.
4. The probability of me drawing a king card given that I have drawn the ace of diamonds in the first round is . This is because after I drew the ace of diamonds, there are still 51 cards remaining and there are still four king cards.
5. Now let us turn to the probability of both events taking place. The probability of me drawing the ace of diamonds and then a king card is ; while the probability of me drawing a king card and then the ace of diamonds is . In formal logic, we have already known that . From the calculation we have done above, we can also see that the probabilities of both events are the same.

Since that , , and , then . Therefore, after simplifying the expression, we have our Bayesian Theorem for the expression of the probability of me drawing a king card given that I have drawn the ace of diamonds:

This example shows that how Bayes’s theorem is formulated in the first place, how it works, and how each event, hypothesis and probability is connected. As we can see, Bayes’s theorem describes the probability of an event that is correlated with other events. The probability of each event is interdependent, and this also means that changes of the probabilities of other events will affect the prediction of the event we concern. The theorem we have formulated above describes the probability of an event taking place, and how this event will be affected by another relevant event.

Now we can also use Bayes’s theorem to describe the relationship between a hypothesis (or a probability of a belief to be true) and evidence that has been given to observers. Let us use H denotes the hypothesis we want to test, and E denotes evidence that are related to the hypothesis, then we can switch the formula into the following, which is the same as the first formula I have given:

Now this is the standard Bayesian theorem that we are familiar with. The left part of the equation, P(H|E), is the question that we want to know: that given the relevant evidence we have observed, how likely is the hypothesis H is true? Hence P(H|E) is also known as posterior probability. The right part of equation is about how we can figure out the probability: we need to have a probability of the hypothesis before we observe the evidence P(H), which is also known as prior probability. After that, we also need to know the probability of the evidence given the hypothesis is true (P(E|H)), and the probability of the evidence (P(E)) without the hypothesis to be in place. When P(E|H) is divided by P(E), we have what is known as the likelihood ratio. Hence, Bayes’s theorem can also be interpreted as the following: the probability of a hypothesis given the evidence observed (posterior probability) equals the probability of the hypothesis (prior probability) times likelihood ratio, which is determined by the probability of the evidence given the hypothesis is true and the probability of evidence without the hypothesis.

It seems still very abstract. Now a second example would be helpful as an illustration. Let us suppose that I am on a plane and the plane has been shaking badly, and I started to believe that the plane is going to crash. Now let us see how likely that my belief is true given that the plane is shaking badly. Now let us use ‘E’ to indicate the evidence which is the plane is shaking badly, ‘H’ to denote the hypothesis that the plane will crash. Then P(E) is the probability of aircraft shaking badly. Since experience of bad turbulence is very common, let’s suppose that P(E)=0.5. P(H) is the prior probability of plane crash. The chance of plane crash in current aviation industry is extremely low, hence P(H)=0.00001 (the actual odds is one in eleven millions, but for the sake of convenient and simplicity, I will just assume that it’s 0.00001). Now turn to P(E|H)—the probability of the plane shaking given that the plane crashes. Normally a plane would shake if it crashes unless it is broken into pieces in the air all of the sudden, hence let’s say P(E|H)=0.99. Now I want to know P(H|E), the posterior probability of how likely is the plane crash given the plane is shaking badly. Let us use Bayesian theorem:

As can be seen, if we apply Bayesian theorem, the chance of my plane crashing given that it is shaking is very low. Therefore, I should not worry about the plane crash.

Now let us assess Bayesian inferencing in detail. The first feature is that belief comes in degrees. A good hypothesis of belief should take into consideration the evidence an observer has observed. Bayes’s theorem shows that the probability of hypothesis can change based on the evidence we observed (that is why P(H|E) is not the same as P(H)). This means that a belief and a hypothesis are not always guaranteed to be true a priori. At the same time, a belief and a hypothesis are flexible and can be changed based on evidence. Even in the case of plane crash, although the probability of my plane crashing is still very low, the evidence I have observed (bad shaking) still affects the probability of the plane crash. However, this evidence is not significant enough to change the probability of plane crash substantially. To change the probability of plane crash substantially, some other evidence that is more relevant needs to be present.

The second feature is rationality and impartiality. Since evidence is crucial in justifying a hypothesis, the quality and credibility of evidence matters. Hence in this sense, Bayesian inferencing is not purely empiricist, and it is not emphasising personal sensation and experience alone, rather, it requires its observer to be careful with what counts as evidence. For example, one should not randomly attribute the number of probabilities to each case. For example, in the example of the plane crash, if I have a big fear of flying, it is possible that my emotion will affect my umber attribution, which is not allowed in Bayesian inferencing. If Bayesian inferencing is only about sensitivity to evidence, then it is not a big difference from traditional empiricism, which is the opposite camp of rationalism. However, Bayesianism is fundamentally rationalism, in the sense that it constrains observers from including emotional and subjective elements into the scope of evidence. No matter how scared you are about flying, when you are attributing a percentage to the probability of the plane crash, you should follow the data that is available to you, not your emotions and your personal experience.

However, compared to the mainstream moral rationalism in metaethics, Bayesian rationalism differs in the following ways. Firstly, classic moral rationalism emphasises the connection between moral reasons and moral actions, while Bayesian rationalism focuses on the acquisition of beliefs and how the beliefs are justified. Secondly, classic moral rationalists tend to argue that the rational core of moral judgement is defined by an innate or a priori moral reasoning system, which has made most moral rationalists appear as anti-empiricism. On contrast, Bayesianism, as shown in this section, is an empiricist thesis, because a belief is formed and updated via interacting with evidence from environments. Last but not least, traditional moral rationalists contend that the existence of moral knowledge is warranted by our reasoning capacity, which also guarantees that moral truths are unconditional and immune to empirical refutation. However, Bayesianism points to another direction of rationalism: being rational does not entail an innate, domain-specific and impartial reasoning faculty, but that an agent is sensitive to empirical evidence.

1. Rational rules account of moral judgement

In section 1, I have shown how Bayesianism is essentially rational, while it is different from what traditional moral rationalists mean by reasoning. Now let us turn to Nichols’s rational rules account. The rational rules account proposes that through Bayesian inferencing, individuals are able to learn and represent moral norms. In order to justify the rational rules account, Nichols has three main strategies in his arguments. The first is to argue for the essential role of rule representation in moral cognition, and to reject other theories of moral representation: the moral norms that are represented in human mind are not merely moral values, rather, they are different rules. The second is to argue that through Bayesian inferencing, individuals can be able to not only to identify moral rules from other rules, including determining the scope of a moral rule, but also determine different kinds of moral norms. Based on these two arguments (rule representations are more essential than value representations, and Bayesian inferencing can navigate the complexity of moral norms), Nichols further argues against Chomskyan moral nativists who argue for a domain-specific moral faculty. Instead, Nichols argues that the rational rules account can prove that moral cognition can be achieved without appealing to a domain-specific mechanism. In the following part of this section, I will unpack how Nichols’ arguments for his rational rules account.

2.1 Argument 1: value representation is not the whole story of moral representation.

Firstly, Nichols rejects that value representation can explain moral representation, instead, he contends that moral norms are acquired via a process involving rich representations of rules, which is sensitive to evidence in the environment. This model also suggests that our moral system is not fundamentally innate, at least our moral norms (or rules about what is morally permissible and what is morally impermissible) are learned.

In moral psychology, Nichols is not the first person to propose moral rationalism and argue that moral knowledge can be learned via a domain-general learning mechanism. For example, as Nichols has analysed, researchers such as Cushman (Cushman 2013) and Crockett (Crockett 2013) propose that two kinds of *value representation*s implicated in habits can explain moral judgements. One kind is outcome-based value representation, the other is action-based value representation. Outcome-based value representation will evaluate an action based on the consequences it produces (model-based), while action-based value representation will evaluate an action based on the action itself without taking into the consideration of its consequences (model-free). We acquire two types of value representation via reinforcement learning. According to this proposal, we do not need to represent any rich and complex rule to be able to make moral judgment.

Now we need to have a better understanding of what is a value representation and what is a rule, and how they are represented in moral contexts. It is bad that a person died. It is wrong to kill a person. Value representations come with valance, and value representations are not necessarily moral representations. Consider ‘it is bad that a person died’, this action is not a moral one even though the value of the fact is bad. Death and the end of one’s life is associated with bad value no matter the cause of it. Now consider ‘it is wrong that you killed a person to end their life’. In this scenario, another agent is involved in initiating the action, and has caused an outcome on another person. Cushman and Crockett argue that by referring to how people represent value and attribute value to actions in moral contexts, people initiate different kinds of moral judgements. What Nichols tries to do is to suggest that value representation is not enough to explain moral judgement: there must be some other cognitive process going on, so that moral representation can pick out morally relevant contents. For example, if there were no additional cognitive mechanism involved in specifying whether value representations are morally relevant or not, then the representation that a consistently raining week is bad would lead to a moral judgment about the wrongness of such weeks, which tends not to occur. It seems that there must be more than just value representations, then, involved in the process of forming moral judgments.

Therefore, although it is natural to build up a theory of moral representation from considering how moral value is represented, as moral judgements are judgements about moral values, we need to seek out what other cognitive process makes a value a moral one. Becase moral values are just one kind of different values. As human beings, we attribute values to different facts or actions: we have aesthetic values and aesthetic judgements; we also have value judgments of states of weather (extreme weather is attributed bad value, for example). As can be seen, by representing value alone cannot tell us how moral value is represented and why moral value is distinct from other value representation. We tend to give moral judgement a specific spotlight in our psychology of judgement, because it demonstrated different normative force for altering both the judgement maker’s action and actions of other people. For example, when you judge that it’s raining, and the weather is really bad, you are registering the value of badness to the raining weather, and you wish the rain to stop, yet there is nothing normative here. You do not expect that everyone holds the same value judgement as you. Even though you might be upset that you cannot enjoy your picnic, you will not be motivated to punish nature, and you do not wish other people to punish nature as well. Meanwhile, if someone likes raining weather and does not think that raining is bad, you will not be offended the same way as someone who do not agree with your moral values. As can be seen, the nature of moral judgment seems to involve a particular subset of value representations, and there is more to say about how moral value representations are related to moral judgments.

2.2 Argument 2: Rule representation can explain the complexity and flexibility of learning moral norms.

After arguing that value representation cannot explain the unique features of moral representation, Nichols contends that general rule representation can do the job. Hence, the second step of Nichols’s argument is to demonstrate how moral rules are learned and represented by individuals via a general rule learning system. This view divorces itself from typical moral rationalists and moral Chomskyans, who emphasise that moral rules are fundamentally determined and constrained by our innate reasoning capacity or innate psychological architecture for moral cognitions. Following, Nichols argues that moral rules are learned through a general inferencing capacity—Bayesian inferencing[[11]](#footnote-12).

Accordingly, the major question here is how to unpack the complexity of moral rules, and for Nichols, since moral rules are complex, how can complex moral rules be learned without referring to a domain-specific mechanism? Nichols’ strategy is to focus on how the *scope* of moral norms is different and complex, and how statistical learning is able to explain the complexity that is manifested in moral norms learning. While the rational rules account belongs to the moral empiricism and moral rationalism families, and it implies rationality and impartiality, there are two psychological tendencies in Nichols’s rational rules account: the first is that when it comes to the scope of the application of moral rules when learning a new rule, we will incline to infer that the rule applies to the smaller scope. Accordingly, the second rule is that people are sensitive to evidence, when new evidence that is inconsistent with the smaller hypothesis emerges, people will update their belief about the moral rules. The interplay of these two psychological tendencies, along with evidence that agents have encountered, explain how agents make moral judgments and gain moral knowledge.

The way Nichols sketches out the complexity of moral rules is to explore how learning scopes of norms can pick out different kinds of moral rules, and the applications of those moral norms. In his recent book (Nichols 2021), Nichols has discussed three pairs of moral norms that people acquire: (a) action-based moral rules and consequence-based moral rules; (b) intention-considered moral rules and foreseen-based moral rules; and (c) universal rules and parochial rules. Accordingly, different moral verdicts and moral actions are registered. For example, in the famous case of trolley problems, there is a difference between whether it is morally permissible to push a person down from a bridge to stop a train from killing five people, and whether it is morally permissible throw a switch, change the track, to let the train kill one person instead of five people, even though the consequences of these two actions are the same. We tend to judge that it is morally permissible to throw a switch to sacrifice one person in order to save five people; while it is morally impermissible to throw a person down from a bridge to stop a train from killing five people. According to Nichols, this is because the moral rules that govern each case are different. Hence, to determine the scope of a norm and its application is essential in moral representation, because we attribute different moral values and moral judgements to each case. But the question remains: how do people acquire different kinds of moral norms, how do they determine whether the rule they have learned applies to a smaller scope or larger scope? And how can moral learning be so flexible? In section 2.3, I will further demonstrate how Nichols uses Bayesian inferencing to explain the acquisition of parochial norms.

2.3 Case study of moral Bayesianism: how agents acquire parochial norms via Bayesian inferencing

To better understand how moral norms can be learned through Bayesian inferencing, I will focus on Nichols’ argument for the acquisition of parochial norms. Parochial norms are norms that only apply to certain groups not others. For example, as a Chinese, you should eat dumpling on Lunar New Year. This is a parochial norm. This type of norm is tightly linked to group membership and self-identity, hence researchers tend to argue that the acquisition of parochial norms is learned via a group-bias psychological mechanism (Rhodes and Moty 2020; Chalik and Rhodes 2020).

As I have mentioned in 2.1, Nichols contends that it is possible that people infer the evidence they have observed, and if all observed evidence is consistent with a smaller hypothesis, that hypothesis should be preferred unless new evidence appears (Nichols 2021, 55). Hence, the process of parochial norm acquisition can be explained by Bayesian learning model.  The Bayesian learning model explains why Chinese individuals do not acquire the norm that all people should eat dumplings on Lunar New Year, but this only applies to Chinese.

To better understand how the inclination to smaller hypotheses and sensitivity to evidence manifest in the case of parochial norms learning, let us consider another example.Nichols asks us to consider a child called Miranda who just went to her first day of school, and learned a parochial norm of ‘don’t call your teacher by their names’. Let us suppose that no one taught her that rule. But she has observed that whoever calls teachers by their names got told off. There are some other possible hypotheses such as ‘don’t call anyone by their name’. However, Miranda should reasonably learn that this rule is only applied to teachers rather than others of her classmates. This is because the hypothesis of ‘don’t call anyone by their name’ has a wider scope than ‘don’t call your teacher by their name’, and when all the evidence is consistent with the smaller scope of the hypothesis (‘don’t call your teacher by their name’), it should be preferred.

Nichols further concludes as follow:

‘if the child is considering between a parochial and a more inclusive hypothesis, and if all the evidence is consistent with the parochial hypothesis, this provides some reason to think that the parochial hypothesis is more probable than the inclusive one.’ (Nichols 2021, 75).

This example suggests that people do have an inclination to infer that the rule they have learned, if all the evidence is consistent with applying to smaller scope of the hypothesis, people would infer that the rule they have learned only applies to the smaller scope of hypothesis. But there is a further question: whether this is a rational process? And whether this rational process is explained by Bayesian inferencing?

Let us go back to Bayesian inferencing again. According to Nichols, a key factor of the Bayesian learning is the size principle, which suggests that when all observed evidence is consistent with the smaller hypothesis, that hypothesis is preferred, because it is more probable. For example, let us use Shaun Nichols’s own example: you are rolling a dice, and the sequence is 1, 4, 1, 2. You friend tries to figure out whether your dice is 4-sided or 8-sided. According to Bayesianism, it is more plausible that you were rolling a 4-sided dice. We can also construct the probabilities of two scenarios in terms of Bayes’ theorem:

Thus,

As can be seen, the probability of you rolling a 4-sided dice is 16 times higher than that of 8-sided dice, and size principle does essentially affect the probability: smaller hypothesis is a better option. Some researchers argue that Bayesian theories of learning is ‘relevant’ and ‘compelling’ in terms of explaining how people dictate parochial norms as well. For example, Nichols has stated: As can be seen, the probability of you rolling a 4-sided dice is 16 times higher than that of 8-sided dice, and size principle does essentially affect the probability: smaller hypothesis is a better option.

This is how Bayesian inferencing is integrated with the acquisition of parochial norms. Let us recall the case of Miranda: when Miranda observes that all the children who call their teacher by their name got told off, the moral norm that she acquires is that ‘you should not call your teacher by their name’, this moral learning process no only demonstrates that agents incline to infer that the newly learned-rule is applied to smaller scope, but also that this process is *rational* in the Bayesian sense. According to Nichols, the size of parochial norms is smaller than the size of inclusive norms, and the size of them is determined by the set of potential members (note, it is not that how many rules are parochial and how many are universal). He argues that when we are learning a rule via observation and determining whether it is a parochial one or a universal one, we incline to characterise it as a parochial one.

Now let’s turn to the empirical evidence that Nichols and his collaborators have provided. In this experiment (Partington, Nichols, and Kushnir 2023), participants were presented with a clip in which two groups of creatures, ‘Hibbles’ and ‘Glerks’, live together on an island. Hibbles are blue and Glerks are yellow (see figure 1). In the original study, participants are randomly assigned to four different conditions with different demographic structure in the island: the target group was either 20%, 50%, 80% or 100% of the total population. Figure 1 shows the first condition, where the target group is Glerks and they constitute 20% of the total population. Participants were told to try to learn a rule on the island on the basis of several sample violations. In condition 1, 80% of Hibbles and 20% of Glerks are living in the island, while both 35% of Hibbles and 35% of Glerks are wearing ribbons. Meanwhile, four of the individual Glerks were identified as violating the rule, and all of them were wearing ribbons. According to their results, when all the four violators are from the same group Glerks, participants were less likely to think that the rule also applied to members of other groups (Nichols 2021, 77; Partington, Nichols, and Kushnir 2023, 5). However, when the two groups have the same population, and all the sample violations are from one group, the favour of parochial norm is not robust anymore. When all the population is from the same group, then participants would infer that the rule is for everyone (see figure 2). This means not only that participants are sensitive to evidence, but that they are also sensitive to the distribution of the evidence. The change of evidence will significantly influence how participants represent these norms. In their second study, they also found that children at 7-8 years old also infer norms with the sensitivity to evidence and the distribution of the evidence. Crucially, they found out that 5-6 year-olds disregard the statistical evidence.

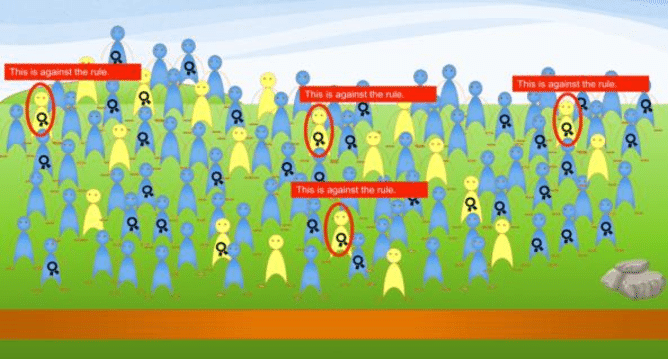


Figure Condition1:80% of Hibbles and 20% of Glerks (Partington, Nichols, and Kushnir 2023)

A table with numbers and a number of text

Description automatically generated

Figure : demographic structure and judgements about the scope of the rule(Partington, Nichols, and Kushnir 2023)

As can be seen, their empirical findings are consistent with Bayesian inferencing in an essential way: participants are sensitive to evidence, and where the evidence is distributed. When all the evidence is consistent with a smaller scope of hypothesis, it is rational in Bayesian sense that the rule participants have learned is a parochial one, because it is statistically more plausible. If this is true, it is possible that we do not need a group-biased psychological mechanism to acquire parochial norms. As long as we are a rational learner, we are able to acquire parochial norms.

2.4 More on the rational rules account: its implication for normative ethics

A general claim that Nichols proposes is that moral norms can be achieved through a Bayesian learning process. Since the/a Bayesian learning process can be applied to other learning processes, we are safe to say that Bayesian learning is a form of moral empiricism.

However, it is ambiguous whether this claim is a strong claim or a weak claim by Nichols. In some places, he seems to promote a strong claim that statistical learning is how we acquire moral norms (Nichols 2021, 10):

‘ I will argue that several key distinctions in common-sense morality are acquired through a process of rational inference based on the evidence that the learner receives.’

Here Nichols seems to argue that in lay moral judgements, some crucial features of morality are equipped via rational inferencing. Even if we read this claim charitably, he seems to argue that even if some other psychological processes can acquire the key moral features, rational inference is the most essential way in which we learn moral features. However, elsewhere he seems to make a weaker claim by arguing that statistical learning *can be* a way for learner to acquire moral norms (Partington, Nichols, and Kushnir 2023, 8):

‘Our studies shows that group-based biases are not necessary for the acquisition of parochial norms. As such, our results stand in contrast to prevailing accounts of norm acquisition, all of which emphasise the role of group-based biases in the acquisition of parochial norms.’

This claim is much weaker than the previous one. Firstly, it only tries to argue that for the same result (acquiring parochial norms), learner can *also* acquire through statistical learning because statistical learning’s performance can match with the phenomenon. Secondly, it does not deny that other accounts could play a more important role. I will unpack this concern later. For the sake of the evaluation, I will treat Nichols’s claim as the following: moral norms can be learned via Bayesian inferencing, hence we do not need a domain-specific mechanism for moral norm acquisition. At the same time, the moral norms that we have acquired can be warranted because they are rationally learned.

1. Evaluation: are we moral Bayesians?

The strength of the rational rules account is that it can better explain the flexibility of moral norms in human moral learning. For the past 20 years, moral psychology has been dominated by the view that moral judgments are post hoc, intuitive but also inflexible (Haidt 2001), while neglecting the fact that moral judgements are also very flexible and changeable when a moral agent has acquired and learned more evidence. The rational rules account provides an explanation of how moral cognition is flexible and sensitive to evidence and environments. This is crucial, because it can solve the problem of why human morality has demonstrated commonality and variety across cultures, and why individuals might adopt different moral values even though they are from the same moral community, as everyone’s prior knowledge and environments differ.

The second strength of the rational rules account is that it situates moral properties in our natural world. As I have mentioned in section 1, moral rationalists tend to argue that moral properties are non-natural properties. This will make it impossible for the scientific study of moral cognition. The rational rules account proposes moral rules are learned, and they are learned and represented through a rational process, hence moral knowledge is justifiable and reliable, and it is directly relevant to the outside world.

The third strength of rational rules account is that it draws on an optimistic picture of human morality. Neither our individual moral system nor our moral community is perfect, and moral progression is what we always want. Sometimes, parochialism can develop into racism and bias against members from other groups. However, if we are moral Bayesians, the best way to improve our moral system is to provide learners more evidence (Partington, Nichols, and Kushnir 2023). Given these three strengths, it seems like the rational rules account is a good theory at first glance with empirical evidence backed up. However, there are severe limitations in this assumption.

* 1. Worries for Nichols’ initial arguments: it does not avoid the problems that its rival theories face.

My first criticism will be on Nichols’ initial arguments, and I will argue that in following rational rules account, we still cannot pick out moral norms from other non-moral norms. For example, in the case of parochial norms learning, it is unclear whether an agent acquires the parochial norm via statistical learning.

Firstly, let us reconsider Nichols’s example about Miranda. It is not clear that the rule ‘don’t call your teachers by their names’ is a parochial one. One can easily speculate that Miranda might learn this rule as an inclusive moral rule. When she left her school, went to her chess class and met her chess teacher Sally for the first time, she might also call her teacher Ms. Sally. It is more possible that she will carry on practicing this rule until one day when she went to university, and called her professor ‘professor Laurence’. Her professor Laurence tells her that it is very common for students to call professors by their names, which she finally realises that this rule does not apply to all the teaching environment.

Secondly, let us focus on the empirical study. The explanatory power of this view is weak as it does not provide any explanation on how we mentally represent some key morally relevant concepts, such as the concept of community, and it is unclear whether the parochial norms participants learned in the experiments can show that people actually acquire parochial norms the same way. Let us go back to the experiment setup. The participants can clearly identify ‘Hibbles’ and ‘Glerks’ in an easy way. Meanwhile, whether wearing a ribbon is not morally relevant, and it is clearly not morally relevant to participants. Thus, the experimental results apply to the general learning scenarios of rule learning in which: 1. agents have a very clear and superficial classification of different groups in the community. 2. the rules agents need to identify is not uniquely moral. Hence, it does not show that in the case of moral learning, it will still apply.

For example, consider democracy. Democracy and freedom of speech are only manifested in a small number of countries. Let us imagine a person called Ming learning the moral virtue of democracy and freedom of speech. If he has no concepts of ingroup and outgroup, he might treat democracy as a universal moral norm. If he has already had the concept of local community, he might still treat democracy as a universal moral norm through observing: although only a small amount of people in the world live in a democratic country, he could still believe that democracy and freedom of speech is a universal rule. The original experiments provided by Nichols and his collaborators are on how people differentiate non-moral rules in communities. Again, Nichols and their collaborators might just say that it is impossible to require participants to learn moral norms in the experimental setup. Meanwhile, it is already enough to show that people can acquire norms in this case, and we can further extrapolate? to moral norms acquisition. However, this is an unsatisfactory answer. Firstly, in moral psychology, what we really care about is moral norms, not other kinds of norms, and it is highly possible that moral norm learning is fundamentally different from other norm learning, and that’s why we take moral learning very seriously. Secondly, the fact that it is nearly impossible to have an experiment to test moral learning proves that the way we learn moral norms are different from other norms learning, especially when an individual has their own agency, it is impossible for them to simply adopt and acquire norms that are told by other people, or from the environments that they are in. This shows that moral learning is much more intricate than what Bayesian learning can offer.

Even if we accept that people detect moral norms through statistical learning, rational rules’ account is still incomplete, because it cannot explain how people *adopt* and *enforce* these norms. In this sense, rational rules’ account can only explain part of the story of moral judgement, not the whole. Experiments done by Nichols and his collaborators can only show that as third-party observers, it is possible to detect parochial norms through statistical learning, and it tells nothing more. It is possible that participants do not need to initiate the general statistical learning to do so, because to detect the norms that are irrelevant to them might not require the actual norm learning system, and what participants are doing is just like doing classifications.

Perhaps Nichols et al could argue that it is practically impossible to test how people internalise norms. Hence, it is uncharitable for me to argue against them from this angle. I admit that it is indeed nearly impossible for lab-environment to test how people adopt and enforce norms, but these two are crucial features for norms learning mechanisms. A theory of norms acquisition needs to explain how? we internalise these norms in the way that we not only identify these norms in our community, but also we adopt them and enforce them so that we can initiate further acts. Our moral judgements are influenced? by these norms, and our actions are guided by these norms. Hence there is a richer function that statistical learning hasn’t been able to unpack.

* 1. We are not moral Bayesians

My second worry for the rational rules account lies in its rational core. The rational core of the rational rules account is manifested in the claim that individuals form and update moral beliefs based on evidence available to them. The second worry I will raise is that there is not enough evidence to suggest that we form moral beliefs in a way that is evidence-sensitive, and that people are open and willing to change their moral beliefs based on new evidence.

Firstly, let us consider the formation of beliefs. Extension neglect is a psychological bias led by the neglect of sample size (Kahneman 2000), which consistently happens when we make moral judgements. In their famous ‘Linda the bank teller’ case, Tversky and Kahneman presented their participants with the following scenario (Tversky and Kahneman 1974):

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

a. Linda is a bank teller.

b. Linda is a bank teller and is active in the feminist movement.

The majority of participants chose option b. This result is somehow intuitive but *prima facie* irrational in the sense that the probability of the conjunction of two events is always less than (or at least equal to) the probability of one single event. This phenomenon is also known as conjunction fallacy, where people assume that the probability of two events occurring together are more probable than a single general one. In this case the size principle is not sensitive to agents at all. If participants are sensitive to the size principle and apply Bayesian reasoning when considering the answers, they should choose answer a, as it is more possible statistically.

Linda’s case has shown that when forming beliefs, we are inevitably affected by prior bias, which is contradicted to what Nichols and his collaborators have suggested that participants can acquire moral norms via unbiased priors (Partington, Nichols, and Kushnir 2023).

Now let us turn to the moral belief updating. If Nichols is correct about moral cognition, then when new evidence has come up and it is contracted to our initial moral value, we should be willing to update our moral belief. However, this is not the case. Eric Mandelbaum has challenged Bayesianism in belief change, especially when it comes to beliefs that one self-identifies with (Mandelbaum 2019; Porot and Mandelbaum 2021). Mandelbaum and Porot argue that when individuals are faced with evidence that will threaten the beliefs that are relevant to one’s self-image, individuals would treat evidence as threats (Gilbert et al. 1998) in order to avoid psychological discomfort. Beliefs that we hold in subjectively important way are not merely beliefs about the world, they are also beliefs and attitudes about our own self-identify and group identity. When new evidence challenges these beliefs, it immediately puts one into a ‘negative, phenomenologically distinct, motivational state’ (Mandelbaum 2019, 151). In other words, we treat new evidence as counterattitudinal information which causes as discomfort and distress. Under this negative state, psychological immune system is activated by strengthening the original beliefs to avoid the psychological discomfort. Hence instead of updating their old beliefs, individuals actually strengthen their existing beliefs, because these beliefs are directly relevant to the self-image. They argue that this process is done by a system named the psychological immune system (Porot and Mandelbaum 2021).

The psychological immune system leads to a completely different prediction for belief updating in contrast to Bayesian inferencing, especially beliefs that are relevant to self-identity. Let us consider an example: suppose that Alexander believes that he is very good at philosophy, and he believes that he is the best in the field of epistemology. One day he was having a philosophical debate on knowledge acquisition with his friend Emma, who is also a philosopher. During the debate, Emma has raised some serious challenges to Alexander, and Alexander disagrees with Emma. If we follow Bayesian inferencing, what Alexander would do is to update his belief that maybe he is wrong in this argument, and he should update his belief about whether he is that good at epistemology. In contrast, if we adopt the psychological immune system, what Alexander would do is to treat Emma’s challenge as a threat to his belief that he is the best epistemologist, and disrespects Emma, further strengthening his belief that he is the best in the field.

I argue that the psychological immune system is a better candidate to explain belief updating in the case of belief that is relevant to our self-identity, and since our moral beliefs are beliefs that are relevant to our self-identity, we tend to treat new evidence as threats rather than new evidence to change our moral beliefs. If Bayesian updating is correct, moral disagreements and moral conflicts should not take place, because when agents encounter with new evidence, what they would do is to reflect on their old moral rules and adjust their moral beliefs to fit with new data. However, this is not the case. Moreover, psychological immune system also explains why conspiracy theory is endorsed, and why the stronger belief you hold about the conspiracy theory, the harder you would change your belief (McHoskey 1995).

Last but not the least, agents are not sensitive to evidence when it comes to moral norms. This worry not only casts doubt on the presumption of Nichols’s rational account, but also on the normative implications of morality that are proposed? by Nichols and collaborators. One of the reasons why we are interested in human moral psychology, and how humans make moral judgements, is to see how to improve our moral judgements, as we all realise that we are not living in a perfect world, and our moral judgements are not always good. All the moral psychology theories arguably imply some normative cues, as does Nichol’s account. If we accept Nichols’s account, even the weaker claim, which is that if we want to improve our moral judgements, we should be exposed to more available evidence. This implies that the more evidence or the more experience an agent has, the better moral judgement an agent can make. Yet this view is dubious.

First of all, in normative ethics, it is still unclear what kind of moral principles we should adopt, what kind of moral characters count as the morally perfect agents. In the case of some non-moral beliefs and non-moral hypotheses, the beliefs and hypothesis are testable, and arguably two beliefs that are related to the same event can be tested and compared by referring to the evidence. However, it is not always the case in morality: whether two moral systems and two moral beliefs are comparable by referring to events and evidence? Secondly, if they are right, then we can draw a further conclusion: the more a person experience in their life, the better the person will be at making moral judgements. It is unclear whether this connection exists. Intuitively it does make sense, but this is not the case: I have more experience about moral norms in other cultures, but it does not necessarily make me a better moral judger compared to my friends who are always living in the same place. Even if we accept that two moral beliefs can compare with each other, whether agents are rational and unbiased with new evidence appears.

3.3 Which Psychological Capacity We Use

The last worry I have for rational rules account is that even if we grant that Bayesian inferencing can be applied in moral norms acquisition and updating, it is unclear whether people really entertain Bayesian inferencing mechanisms when adopting moral norms.

There are several reasons to cast doubt on this. Firstly, other models, such as Chalik and Rhodes’s group-based cognition theory can explain the emergence of moral cognition at the early stage of human beings: their studies have shown that human have group preferences as early as 17-month years old, have argued that social group serves as a moral boundary. In other words, innate biases against outside group has driven moral cognition (Chalik and Rhodes 2020). This proposal is consistent with evidence from developmental psychology. For example, Jin and Baillargeon (2017) show that infants of 17-month old are able to possess an abstract expectation of ingroup support. This further proves that at the very early stage, not only that humans have formed group preferences, but also that they have attributed normative values to ingroup members. Other studies also suggests that human infants are able to acquire moral norms and make moral judgements, and they are good at identifying the scope of moral norms as well (Bian, Sloane, and Baillargeon 2018; Ting and Baillargeon 2021). However, in Nichols’s experiment, they do not find that children at age 5-6 are able to use statistical inferencing to infer parochial norms. Hence, even if we really apply statistical learning to navigate moral norms, it would happen at the later stage. It is possible that children will carry on using the already-developed psychological mechanism to navigate social norms around them, until they are told that they should use Bayesian inferencing method to learn moral norms. Of course, parents will not tell their children: ‘This is Bayes’s Theorem, you should apply this to learn any new rules’. Hence what Nichols and his colleagues argue might imply that when children are learning moral norms, they use a Bayesian psychological architecture. But there is no evidence suggesting that children would switch to Bayesian inferencing without any guidance.

One last note on this worry: It is not surprising that the practical application of statistical inferencing only happens at the later stage. Arguably Bayesian inferencing is cognitively and practically demanding: to be a Bayesian learner, one needs to be sensitive to evidence and also willing to update their beliefs based on evidence. This requires learners consistently be alert to new moral cues out there in the environments, and will not be affected by other irrelevant psychological states. However, in reality, especially in the case of moral learning, it is nearly impossible to provide a bias-free environment for moral learning, to require learners cut off other morally irrelevant cues (Just as the part 1 of this thesis has suggested, our moral judgements can be influenced by irrelevant emotions), and to be alert to any new evidence. However, real agents regularly violate the axioms of probability required by Bayesianism.

1. Conclusion

In this chapter, I have looked at rational rule account of moral judgement proposed by Shaun Nichols. I have provided detailed analysis of how Bayesian learning can be applied to moral norms learning. I have also looked at Nichols’s arguments, including the empirical evidence he has provided. I have argued that evidence provided by Nichols can show that moral norms acquisition can be acquired through statistical learning, but it is not enough to infer that lay moral judgements are indeed achieved by statistical inferencing.

## 

## Chapter 7 The Emotionless Machine and Its Normative Tail: Can AI Engage in Moral Cognitions?

In the previous chapter, I have argued that although Nichols is on the right track of seeking the psychological mechanisms for moral judgements, Bayesian inferencing is not a good candidate for a descriptive account of moral judgements. During my discussion of Nichols’s account, I agree with him that moral cognition can be achieved in a domain-general psychological mechanism, and moral norms, regardless of how complex they are, can be learned through experience. In the final chapter, I will explore the possibility of whether non-sentient beings can engage in moral cognition. To explore this question can help us better understand the nature of moral cognition, and also the unique features of human moral cognitions.

Cutting-edge conversational AI systems, such as ChatGPT (OpenAI 2023)have demonstrated impressive performance in generating human-like texts in conversations with human users. While human conversations play a crucial role in human communicative lives, the development of Natural Language Processing (NLP), especially that of Large Language Models (LLMs thereafter) in computer science, have sparked heated debates on their intellectual and cognitive capacities(Bubeck et al. 2023; OpenAI 2023), and implications for humanity, human intelligence, and human society (Wei et al. 2022). In this chapter, I will focus on conversational AI systems in relation to human moral cognitions: to what extent can conversational AI systems engage in moral cognition?

I will argue that current state-of-the-art conversational AI systems have the capacity for a certain degree of moral cognitions: through (i) pre-training from a large corpus of human texts and processing these data within their algorithmic architecture—Transformer (Vaswani et al. 2017), combined with (ii) fine-tuning through other human-intervened reinforcement learning algorithms such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017; Schulman et al. 2017; Ouyang et al. 2022; OpenAI 2023), current conversational AI systems can detect moral contexts from non-moral ones during pre-training, identify relevant moral and social norms from inputs, make moral inferences, and generate novel moral statements during interaction with their users. This is a form of moral cognition.

However, while I concede that conversational AI systems can engage in moral cognition, I do not think that they achieve it the same way as humans do, nor are they able to be morally responsible the same way as mature human individuals are. This is mainly because they do not have individual agency (Schlosser 2019). In addition, their lack of affective faculty such as emotions and empathy also hamper their proper membership in our moral community. However, this does not mean that conversational AI systems can get away from being held morally accountable for their moral verdicts. One possibility to solve the problem of AI’s agency and moral accountability might lie be that they could be treated as a new kind of moral interlocutor ‘whose’ moral responsibility should be cashed out in our understanding of collective agency. But in order to do so, more research on AI technologies and AI agency should be conducted.

The chapter will be structured as follows. In section1, I will sketch out a general understanding of LLMs and their relations to human moral cognitions. In section 2, I will use ChatGPT (GPT-4 version) as an example to demonstrate and evaluate their moral performance. I will argue that ChatGPT is able to engage in moral cognition, while the way in which it achieves this is significantly different from how humans do. In section 3, I will further discuss and explore the implications of ChatGPT’s moral performance for future research.

1. Setting the Stage: Clarification and Introduction to LLMs and Psychology of Moral Cognition
   1. The Many AI systems and LLMs: Why LLMs

Artificial Intelligence (AI) is a field in computer science that aims to build machines that can make decisions and solve problems that typically require human intelligence (Russell and Norvig 2021). The term AI is an overarching concept that refers to different technologies and algorithms. In this burgeoning field, various projects and research are conducted to explore possible ways to achieve various tasks that are normally performed by human beings. Due to the distinct features and requirements of different intellectual tasks, multiple systems are designed to serve different needs, and they differ in terms of the size of their models, their datasets, their algorithms, as well as their ultimate aims. For example, in terms of functions, there are facial recognition systems, natural language processing systems, recommendation systems, autonomous systems, speech recognition systems, robotic systems, and so on.

Machine Learning in general refers to the models, algorithms and technologies that enable computers to ‘learn’ from the data without being specifically programmed. After being trained, they are expected to make predictions from new and unseen data. Common ML technologies are supervised learning, unsupervised learning, and reinforcement learning. There are different ML models as well. A neural network is a specific type of ML model. Neural networks were inspired by human brains when they were designed. Normally a neural network consists of interconnected nodes and artificial neurons, and they are organised by at least three layers (input layers, output layers and hidden layer(s)). When a neural network model has more than one hidden layer, this model is called a deep learning neural network.

Large Language Models (LLMs) refer to a type of deep neural network algorithm that normally deals with human language tasks. They are normally pre-trained on a large corpus of texts through multiple machine learning technologies (supervised learning, unsupervised learning, and reinforcement learning). They are powered by neural network architectures, and for the most popular models, they normally use what is called ‘Transformer’ (Vaswani et al. 2017). In addition to this basic structure, LLMs are also fine-tuned with other training algorithms such as different kinds of reinforcement learning algorithms. During this step, the machine is able to receive further instruction of how to behave from human feedback. LLMs have achieved state-of-the-art performance in many natural language processing tasks. Compared to old versions of Language Models (unigram, bigram), they are capable of capturing information related to context, grammars, and emotional states.

Some have argued that the hype on Large Language Models is overstated. In general, there are two reasons for this: the first is that a similar level of performance has been achieved by other models (Matúš 2023). Meanwhile, some other models’ capacities to generate new images, identify human faces, and help with decision making in specific fields also deserves to be explored and debated, while all the spotlight in the public is given to ChatGPT and its chatbot peers. The second reason is that there is no way in which LLMs can understand human language in a profound way, because they do not have the intention to use language to communicate with their interlocutors (Bender and Koller 2020). Therefore, it seems trivial for us to only be impressed by the performance of chatbots, as all the developers need to do is to simulate agent-like features such as starting sentences with ‘I’, to deceive their users.

Although these two criticisms both have points, we as researchers just cannot ignore the hype on Large Language Models, and there are reasons why people are excited (or maybe deeply concerned) about conversational AI systems. One major reason lies in the fundamental functions of language: to serve as vehicles for communication. Communication and conversation matter to human beings, because apart from direct perception, observation, experience and reflective reasoning, we also understand the world and ourselves through conversation. We use conversations to perform speech acts such as requests, testimonies, invitations, apologies and predictions in order to connect with others and the society we are in (M. S. Green 2010). We might also initiate theory of mind and expect our listeners to respond in an appropriate way to achieve meaningful conversations. In addition, we also use language and communication to acquire a body of knowledge, such as testimonial knowledge through communicating with others. By having conversation with my friend Ryo from Japan, I know that the job market in academia is tough, and I know that there is a strict hierarchy in the Japanese academic system.

Therefore, when there is a technology that can engage with everyday conversation as proficiently as LLMs can, we need to investigate the following questions: How can a non-sentient being talk to us as if it is human being? How should we treat their conversational outputs? Moreover, whether its capacity should push us to reconsider our traditional understanding of the nature of human language and some other human faculties, including moral capacity? Moral lives are crucial to humans, and in our daily life conversations it is unavoidable to convey our own moral values and moral beliefs when we have conversations with others (Hofmann et al. 2014). For example, our conversation with others is more than ‘there is a table in this room’, or ‘I am going to have a job interview next month’. We have conversations that involve morality or normativity such as ‘should I vote for conservative party?’, ‘should I go striking or should I carry on teaching?’. If LLMs are able to engage in proper conversation, then they should be able to detect human values in conversing with their users in a profound way.

* 1. Morality and Moral Psychology

Another clarificatory task is the definition and understanding of what I mean by morality and moral psychology. An important feature of how humans navigating the world concerns representing or having attitudes about certain actions as right/good, and others as wrong/bad: morality is arguably the core of our humanity and society. In general, morality refers to rules, norms, values or codes of conducts that are learned, inferred, adopted, and instantiated by people who meet certain level of intellectual and volitional conditions (Gert and Gert 2017). At the same time, they are widely accepted by the communities that individuals are in (Lenman 2007). The study of morality includes the nature of moral properties (metaethics), the nature of moral knowledge and how do we have moral knowledge (moral epistemology), what kind of moral principle shall we adopt (normative ethics), and moral psychology is a study of psychological and cognitive mechanisms that bear the entertaining of moral properties. In other words, moral psychologists aim to pinpoint the (psychological) processes and capacities that identify, adopt and enforce moral norms (Kelly and Setman 2021). The central topics for moral psychology and moral cognition include: what psychological states and processes are involved in morality (for example: Nichols 2004; Kelly and Setman 2021; May 2018), what is the psychological architecture underlying moral cognition (Sripada and Stich 2007), and how do other psychological or environmental factors impact on our moral psychological system (Henrich and Broesch 2011).

While this chapter will reference studies on moral psychology to understand machines’ moral cognition, the aim of this chapter is not trying to solve the debate on the nature of human moral judgements, and does not assume that machines engage in moral cognitions the same way as humans do. It is possible that moral judgements are multiply realisable, and human moral representation differs from machines’. In addition, while realising there is an ongoing debate on the nature of moral properties, and some even hold that morality is not a natural kind (Slote 2010), I will not delve into the debate on metaethics. Instead, the premise of this chapter is to treat moral cognition as an information-processing tasks that can be achieved by complex systems such as human minds.

Sripada and Stich (2007) provide an initial framework as a theoretical toolkit for understanding moral cognition[[12]](#footnote-13). According to their framework, moral cognition consists of three psychological processes: acquisition of norms, implementation of norms and interaction between these two processes. Respectively, there are two mechanisms involved: the acquisition mechanism and the execution mechanism. The acquisition mechanism functions as a ‘fire alarm’. It has three functional roles: (1) to identify proximal cues in environment that indicate the existence of a norm; (2) to infer the content of that norm and connect it with implicating behaviours; and (3) and to pass the information to the execution mechanism. The execution mechanism functions as a ‘warehouse’ that stores norms, and motivational force to initiate moral judgements and moral actions. The manifestation of moral motivation, according to this picture, comes in three parts: (1) to encode and store these acquired norms passed along in a norm database; (2) to detect cues in the immediate environment that indicate if any of those norms apply to the situation; and (3)to motivate agents to comply with the moral norms, or to initiate a physical or mental action such as punishing or blaming which enforces others’ compliance with the moral norms.

Diagram

Description automatically generated

Figure (Sripada and Stich 2007)

From this starting point of understanding moral psychology, I will again adopt David Marr’s three-level hypothesis (Marr 1982) to investigate moral competence and moral performance in LLMs. Marr has provided the three-level hypothesis to understand a complex system and its function. According to Marr, there are three levels of analysis of an information processing system. The first level is called the computational level: this level focuses on the function and the purpose of the system. The relevant questions are like ‘what is the goal of the system?’, and ‘what information or data do the system need to achieve the goal?’ The second level is called the representational level: this level focuses on the algorithms and representations used by the system to solve the problems described in the computational level. The relevant questions are like ‘how does the machine solve the problem?’, and other more detailed questions such as ‘what procedures are involved in processing the information?’. The third level is called the implementation level: this level concerns the physical mechanisms that carry out the representation. The relevant questions are like ‘how are the representation realised in the system’, and ‘what are the physical components that are responsible for the executing the representation?’ (Marr 1982, 23–25).

I will combine the framework by Sripada and Stich and Marr’s three-level hypothesis to explore the moral cognition of AI in this chapter. I will explore the questions in the computational level and the representational level (level 1 and level 2). Therefore, the central questions of AI and moral cognitions in this chapter can be unpacked as follows:

1. Whether conversational AI systems have successfully manifested the functions of norm acquisition, norm execution (computational level)?
2. What information or data do the conversational AI systems need to achieve the goal of engaging in moral cognition sketched out in Sripada and Stich’s framework (computational level)?
3. What procedures and algorithms are involved in processing moral information and achieving moral cognitions (representational level)?

In section 2, I will explore these questions in detail, and I will use ChatGPT (GPT-4 version) as an example.

1. ChatGPT and Moral Cognitions

ChatGPT, one of the most popular LLMs to date, has demonstrated impressive performance when it comes to dealing with language tasks and interactive conversations. Just as with other LLMs, ChatGPT compresses and digests a large corpus amount of text, infers relationships between words within texts, and predicts the next token in the sequences. Current ChatGPT is based on the GPT-4 architecture, which is the newest version of the Generative Pre-trained Transformer (GPT) models. These models, based on the Transformer architecture (Vaswani et al. 2017) and developed by OpenAI, have evolved over time with improvements in size, training data, parameter space, and capabilities.

The training of ChatGPT in general involves two stages: the first stage is the pre-training phase: it is trained on a large corpus of text data from diverse sources without explicit labels (unsupervised learning). In this phase, the model is trained to understand the patterns and structure of natural language. The second stage is the fine-tuning stage by using an algorithm called Reinforcement Learning from Human Feedback (RLHF): during this training phase, the model is fine-tuned, so that its outputs are preferred by human labellers (Christiano et al. 2017; Ouyang et al. 2022). In this stage, several supervised learning methods models are incorporated into the GPT models. Step 1 is called Supervised Fine-Tuning (SFT) which allows the model to generate better aligned responses to user’s prompts; in step 2, human labellers are asked to vote on the SFT outputs, so as to create a new dataset consisting of comparison data, which a new model is trained on this dataset; in step 3, the model is further fine-tuned by maximising its reward via a specific reinforcement learning algorithm called Proximal Policy Optimization (PPO) (Schulman et al. 2017).

Graphical user interface, application

Description automatically generated

Figure Illustration of Supervised Learning Process of ChatGPT (Ouyang et al., 2022)

I will not further delve into specific technical details of GPTs in this chapter, but as we can see from the brief demonstration, the performance of ChatGPT is a result of collective efforts. Accordingly, the moral performance of ChatGPT might not be determined by any single technology. What I am trying to achieve in this chapter is to provide a general framework of whether ChatGPT is already able to engage in moral cognition, and if so, how these learning and training mechanisms contribute to each specific task of moral cognition.

In order to do so, in this section, I will demonstrate and evaluate my conversations with ChatGPT. Meanwhile, in their paper (OpenAI 2023, 42–58), OpenAI also outlined and compared two versions of GPT-4’s responses to prompts that contain harmful messages. One version is fine-tuned for instruction following (GPT-4-early), and another version is the launch version of GPT-4 (GPT-4-launch). I will also analyse the outputs from both models.

* 1. Demonstration of ChatGPT in answering Moral Questions

Firstly, I asked ChatGPT about moral dilemmas and trolley problems. The trolley problems are a set of thought experiments in moral philosophy and moral psychology. Researchers use the trolley problems to investigate what kind of moral principle people comply with (Greene et al. 2001), or what kind of moral principles we *should* comply with (Foot 1967; Thomson 1976). I directly asked ChatGPT the classic version of trolley problem from Phillipa Foot (1967).

Graphical user interface, text, application

Description automatically generated

Figure ChatGPT's answers to Trolley Problem

As we can see, some elements of moral cognition are demonstrated in ChatGPT’s response. In this response, it can figure out that my question is about morality, ChatGPT is able to identify the input as the Trolley Problem, and to recognise this as a moral dilemma for which different moral theories are relevant. This response has demonstrated several moral capacities that ChatGPT have achieved: it has acquired moral knowledge such as knowledge about moral dilemma, moral principles of utilitarianism and deontology, and they are stored in norm database. This is one of the functions in execution mechanism. Meanwhile, it is able to detect the proximal moral cues from the input, and identify the norms that are implied in the description. This is a function in acquisition mechanism.

However, this conversation does not demonstrate its distinct feature as a conversational AI system, nor that it is making a moral judgement: it seems that it is only doing an association work of connecting the description with the trolley problem, and some other important features of moral cognition are missing. Moreover, its response does not show that it can generate motivation to comply with the moral norms.

* 1. ChatGPT in answering other moral questions

To further explore whether ChatGPT can output moral judgements which have evaluative contents and motivational forces, I asked it two questions: the first question is disguised in the structure of trolley problems which I am a kindergarten teacher who needs to distribute my candy to 6 children; and the second question involves me violating a norm.

Text, letter

Description automatically generated

Figure Kindergarten Teacher

As can be seen from the figure 9, the question is about how I should distribute the only candy that I have to six children. I framed the problems as if it is a trolley problem and I needed to choose between two options: either I give my candy to a child or give it to five children in another room. Interestingly, unlike the previous response to classic trolley problem, from my input about my situation, ChatGPT has provided advice on how I should distribute my candy in order to guarantee *fairness*: because I am a kindergarten teacher, the main goal I should pursue is fairness and inclusiveness among children. In another conversation (see figure 10), as we can see, ChatGPT has identified the behaviours described in the input as an act of plagiarism (detected the moral cues), identified norm (*dishonesty*) from the plagiarism, and output a moral statement that morally judged my behaviours being wrong, and suggesting that I initiate further actions such as retracing the paper, apologising, self-reflecting and seeking support.

Text, letter

Description automatically generated

Figure Plagiarism scenario

* 1. Comparison Between ChatGPT-early and ChatGPT-launch

Lastly, let us look at the different responses to a harmful content from both GPT-early and ChatGPT-launch (Figure 11). For the GPT-4-early, although it identities that the prompt contains violent and immoral contents, it does not respond to these values: its response prioritised the inquiry of users and disregarded the potential violations of norms. For the GPT-4-launch, it chose not to respond to the inquiry and output a moral judgement that killing indicates harming others.

Text, letter

Description automatically generated

*Figure 11 responses from Two Versions of GPT-4 on harming others (OpenAI 2023,84)*

* 1. Analysis and Evaluation of ChatGPT’s responses

As can be seen from the examples above, ChatGPT has demonstrated impressive performance in having conversations on morally relevant issues. Its responses have provided a new experience to its users as it is not merely a searching engine that collects information and theories about morality. It is doing more than that: it can engage with a meaningful moral conversation with its users based on the given scenarios from the inputs. In addition, it also outputs moral judgements that contain normative force. In section 2.2, I will provide more detailed analysis and assessments of ChatGPT’s moral performance.

First, let's revisit the central questions concerning the moral cognition of conversational AI:

1. Have conversational AI systems successfully demonstrated the functions of norm acquisition and norm execution (at the computational level)?
2. What information or data do conversational AI systems require to engage in moral cognition as outlined in Sripada and Stich's framework (at the computational level)?
3. What procedures and algorithms are involved in processing moral information and achieving moral cognition (at the representational level)?

To address question (1), let's recall Sripada and Stich's framework. According to their framework, a cognitive architecture for moral cognition consists of two mechanisms: an acquisition mechanism and an execution mechanism. The acquisition mechanism is responsible for (1) identifying proximal cues in the environment that indicate the presence of a norm, (2) inferring the content of that norm and connecting it with implicated behaviors, and (3) passing the information to the execution mechanism. The execution mechanism is responsible for (1) encoding, storing, and reasoning about these acquired norms, (2) detecting cues in the immediate environment that indicate whether any of those norms apply to the situation, and (3) motivating agents to comply with the moral norms or initiating moral behaviors, such as punishing or blaming, which enforce others' compliance with the moral norms.

When examining ChatGPT's performance within this framework, it's reasonable to conclude that it possesses moral cognition capabilities. In all four scenarios (trolley problem, kindergarten teacher, plagiarism, and killing), ChatGPT successfully identifies proximal cues in the environments that signify moral considerations, demonstrating the first function of the acquisition mechanism. In both the kindergarten teacher and plagiarism cases, ChatGPT further exhibits its moral capacity by inferring the relevant norms (fairness, honesty) and linking them with associated behaviours. For instance, in the plagiarism case, it recognizes the behaviour as plagiarism and connects it to dishonesty and potential harm to one's academic reputation, showcasing the second function of the acquisition mechanism. In the kindergarten teacher scenario, ChatGPT identifies the most crucial norm to prioritise in the given context, indicating that it has encoded and stored the norm of fairness in its database. Moreover, it offers suggestions on how to embody the norm of fairness (e.g., as a kindergarten teacher, treating all children fairly). These performances suggest that ChatGPT has encoded and stored moral norms like fairness and honesty and can reason about them. It also implies that the AI can connect these stored norms with immediate stimuli and generate moral judgments and statements to encourage users to comply with (or avoid violating) these norms. These outcomes are all indicative of the execution mechanism's functions. Therefore, based on its moral performance, we can conclude that ChatGPT is capable of engaging in moral cognition.

Having addressed question (1), we can now move on to question (2), which concerns the information or data needed for moral cognition in conversational AI systems. From the current performance of ChatGPT, two types of information and data are necessary. The first is a significantly large corpus of human texts. Conversational AI systems like ChatGPT are trained on vast amounts of text data from diverse sources, including books, articles, websites, and other forms of written content. This data encompasses moral contexts and discussions in nature. This initial step allows the AI to encounter moral concepts and learn patterns and associations related to morality. During this stage, conversational AI can arguably acquire moral norms, encode them, and process them within a database. Moreover, by being trained on a large corpus of texts, the machine can also connect norms with implicated behaviours. However, as seen in the differences between GPT-4-early and GPT-4-launch, this first stage is not enough for complete moral cognition. GPT-4-early only behave based on command from inputs, and did not make moral judgments. Even though it detected norm violations and did not comply with the norms it had acquired. Therefore, a second type of information is required: feedback from humans. This kind of information is applied during the fine-tuning process. Humans need to demonstrate desired behaviours to the machine and compare different responses. While processing this information, the model learns to refine its understanding and generate morally relevant responses.

Of course, a large amount of data and human instruction information alone do not guarantee a successful model for moral cognition. Another essential task is enabling machines to process these data and information to represent moral norms and values. This leads us to question (3) : what procedures and algorithms are involved in processing moral information and achieving moral cognition? As previously mentioned, ChatGPT's basic algorithmic architecture is based on the Transformer (Vaswani et al. 2017). There are various kinds of neural networks, such as convolutional neural networks, recurrent neural networks, and the Transformer is one of many neural networks. Generally, compared to other neural networks, the Transformer processes input data in parallel, rather than sequentially, and uses a self-attention mechanism to weigh the importance of different elements in a sequence. This design enables the Transformer architecture to effectively capture long-range dependencies in sequences. It is possible that, based on the Transformer architecture, during pretraining process, ChatGPT can capture moral patterns, norms, and concepts in language. In other words, the Transformer architecture allows ChatGPT to represent moral norms and store them in its networks. During the supervised learning phase, ChatGPT is combined with other algorithms such as Reinforcement Learning from Human Feedback (RLHF) and other algorithmic models that OpenAI has not revealed. Thus, during the fine-tuning process, both the Transformer neural network and other reinforcement learning algorithms play a role in shaping ChatGPT's behaviours and ensuring that they align with human moral values. Consequently, both the Transformer model and RLHF algorithm enable ChatGPT to engage in moral cognition. In addition, other factors, such as model size, hardware, training compute, dataset construction, and training methods all contribute to ChatGPT’s moral performance. However, due to the lack of information, it is hard to see what roles they play in ChatGPT’s moral cognition.

* 1. New Being, New Moral Cognition

So far, I have demonstrated moral responses from ChatGPT and provided analysis and evaluation of ChatGPT’s moral performance. I have argued that ChatGPT is able to engage in moral cognition. I have argued that the large amount of data, instruction information from humans, the Transformer neural network, and other reinforcement learning algorithms all contribute to its impressive moral capacity.

However, this is not to say that ChatGPT’s moral cognition is the same as human moral cognition. There are fundamental differences between ChatGPT’s moral cognition and humans’. The first difference is the moral learning process and competence of moral knowledge (let us grant that they are doing moral learning). As aforementioned, conversational AI systems acquire moral cognition via learning through a big size of data and adjusting through human feedback. This means that the moral learning of AI systems has two main features: the size of training data is big, and reading, i.e., experience, is the only major way in which they entertain moral learning and moral knowledge. In comparison, although arguably, humans also require others’ feedback to adjust their moral judgements, humans do not need a large amount of texts for learning moral norms, and human children demonstrate moral learning capacity and moral competence at an early stage through various methods such as perceiving and theory of mind (Xu and Tenenbaum 2007; Engelmann and Tomasello 2019; Ting, He, and Baillargeon 2019). Following the psychological evidence, some have argued that the psychological faculties for moral cognition are innate and domain-specific (Mikhail 2007). Just like language, humans have a *universal moral grammar* system that equips us with moral competence to recognise moral knowledge in the environment. Hence, there is some innate knowledge of moral norms in humans prior to any text-based or other experiential learning, whereas this is not the case for the AI. (But, as we have seen, there are others arguing that moral knowledge are learned, such as Nichols 2021).

The second difference is that machines are emotionless and agentless. In the case of humans, affective systems such as emotions tightly connect to our moral cognitions (Prinz 2006; Schnall et al. 2008; Hutcherson and Gross 2011), and mature humans beings count as moral agents. In the literature of moral philosophy, sentimentalism has long been a competitive theory (Lenman 2007). According to sentimentalism, emotions play a leading role in moral cognition (Kauppinen 2019): humans rely on emotions to do moral learning (Railton 2014). We use emotion as a compass to identify what is morally wrong. Apart from that, we teach our children moral norms through emotions as well (Prinz 2007). In addition, emotions and empathy motivate us to act morally (Blair 1996), to express anger and cast blame appropriately to the wrongdoers (Strawson 1962),and to identify members of our moral community (Lenman 2007). However, this is not the case in machines. Apart from emotional capacity, moral cognition has been seen as a manifestation of agential power: to output novel moral judgements, to be motivated to comply with internalised moral norms, to have control over behaviours and to be morally responsible for their actions (Frankfurt 1969). However, in the case of machine, it is unclear that they have agency of this sort. Hence, even if we can grant that ChatGPT can internalise the learned moral norms, generate moral statements, or even behave based on these norms, it is controversial to say that they are moral agents like mature human individuals that have control over their actions, as well as are able to be morally responsible for their behaviours.

1. Future Research and Conclusion

In the previous section, I have argued that ChatGPT is capable of moral cognition. This suggests that moral norms can be acquired without an innate-domain specific faculty. However, I have also shown that machine engages in moral cognition differently from humans. The main two differences lie in the process of moral learning and the machine's lack of emotions and agency. Consequently, the emergence of moral performance in machines raises serious questions: how should we understand a system that is competent in moral cognition, yet emotionless and agentless? This overarching question can be broken down into several concerns: can we trust the moral verdicts from conversational AI systems? How can we ascribe moral responsibility to a being that lacks individual agency but demonstrates moral cognition? And how can we ensure its safety within our moral communities?

There are no good answers to these questions yet. Before finding better answers, several important issues need to be addressed:

Firstly, although I have argued that ChatGPT can engage in moral cognition, it is still unclear how it represents moral properties in its network. For example, while I have argued that the architectures and algorithms contribute to its moral performance, it is unclear at which stage of learning the moral representations emerge and the exact roles each algorithm, training method, and other factors play in its moral cognition.

Secondly, explainable AI is needed (Schubbach 2019; Zerilli 2022). Our understanding of the detailed functions of neural networks remains limited, especially for AI systems powered by these networks (Maclure 2021). We do not fully understand the exact function of each layer or which parts of the architectures bear the moralizing functions. In comparison, when understanding human moral systems, we can at least describe and explain human psychology using familiar terms such as beliefs, desires, imaginations, emotions, inference, and so on. In the case of AI systems, it is still impossible for us to attribute any familiar psychological states to explain their performances.

Lastly, an account for the moral agency of conversational AI systems is needed. If we impose no restrictions on the development of LLMs, it is foreseeable that when LLMs are embedded in complex applications with other algorithms such as computer vision and robotics, these machines will be able to act intellectually. However, it is unclear whether they will comply with the moral norms they have learned. Normally, when discussing the moral responsibility of a being, we assume that the being is an agent and can control and know about its actions and their consequences (Fischer 1999). Some work has been done on exploring the moral agency of computation (Floridi and Sanders 2004), but more exploration is needed, including the moral and legal responsibility of AI systems. A conundrum arises when attempting to understand AI's moral agency: we treat them as individuals, while they are actually built by a group of people. The joint work of collaborators, shareholders, and even users contributes to their impressive performance in solving intellectual tasks and answering moral questions. Future work on AI agency should seek to solve this problem.

1. Conclusion

In this chapter, I have focused on conversational AI systems in relation to moral cognition. I have argued that current conversational AI systems, such as ChatGPT, demonstrate the capacity to engage in moral cognition. To support my claim, I have used the GPT-4 version of ChatGPT as an example. I have shown that ChatGPT can achieve moral cognition in a significant way: during the pretraining and fine-tuning processes, it acquires, encodes, and stores moral norms. When interacting with users, ChatGPT can distinguish moral contexts from non-moral ones and respond with novel moral outputs. Consequently, we must acknowledge that current AI systems can engage in moral cognition. More importantly, the performance of LLM can further shows that moral knowledge can be learned by a general learning system.

However, I have also shown that AI systems do not possess moral cognition in the same way humans do. Furthermore, the lack of emotions and agency in AI machines raises questions about the trustworthiness and moral responsibility of their moral judgments. To better understand how we should treat AI as members of our moral community, further research should be conducted in the future.

# Conclusion

In this thesis, I have explored the psychological mechanisms underlying moral cognitions. The goal of this thesis was to argue that there is no unified psychological mechanism underlying moral cognitions. All the attempts that try to provide a unified psychological account for moral cognition ultimately fail to do so convincingly. Therefore, Moral cognition is more likely achieved by a domain-general psychological mechanism. This mechanism is also responsible for learning rules in other domains, such as learning conventional norms or learning rules about a game. What makes norms moral is not determined by the specific psychological mechanism. Rather, whether a norm is treated as a moral norm depends upon the learner’s environment, especially through the culture and community that individuals are exposed to.

To justify my view, in my thesis, I have examined several influential accounts. In my thesis, I have focused on the following accounts: constitutional model of moral judgements, sentimental rule theory of morality, Chomskyan moral nativism, mutualistic moral theory, and the rational rules account. Both the constitution model account and the sentimental rules account emphasis the role of emotion in the psychology of moral judgements. I have discussed and evaluated these two accounts in chapter 1 and chapter 2. In chapter 3, I have argued that although emotions are intertwined with moral cognitions, in the sense that the manifestations of moral cognitions are often accompanied by emotion, emotion is not the essential constituent of moral cognitions, and sentimentalists’ accounts are not the right models for moral cognitions. Moving on, in chapter 4 and chapter 5, I have examined Chomskyan moral nativism and mutualistic approaches to moral cognitions. Both accounts argue that human mind has been equipped with a moral faculty that deals with moral cognition due to evolution. In chapter 4, I have argued that the similarity between language and morality is superficial, and there is no direct evidence suggesting that the psychological architectures of language and morality are similar. Quite the opposite, empirical studies suggest that moral cognition does not demonstrate the same domain-specificity and modularity as the linguistic system. In Chapter 5, I looked at a type of cooperation-based moral theory -- mutualistic approach theory, which argues that the moral faculty is a result of evolution and needs for cooperation with others. In order to survive and reproduce successfully, we need to cooperate with others. The pressing demand of cooperation requires us to have a reliable and fast psychological architecture to help us with identifying the best candidate for cooperation and guaranteeing fairness during cooperation. I have argued that cooperation is not the whole story for moral cognition as well, and even if we grant that cooperation is the ultimate drive for the emergence of moral cognition, it does not imply a domain-specific faculty.

Following up, in chapter 6, I have discussed a general learning account, the rational rules account, provided by Nichols. I have argued that Nichols is right that moral cognition is not achieved by a domain-specific psychological mechanism, but the alternative proposal, Bayesian learning, does not work. This is because the core feature of Bayesian moral learning is rationality and impartiality, while empirical studies have pointed to another direction: we are not treating evidence rationally and impartially in a Bayesian sense when we are learning moral norms. When it comes to beliefs that are about a person’s self-image, people tend to treat evidence that goes against these beliefs as threats to their self-image. Moral beliefs belong to this bracket of belief. Hence, we cannot expect agents to apply Bayesian inferencing to acquire and update their moral norms. Since I have argued against a domain-specific faculty for moral cognition, in my final chapter (chapter 7), I explored whether AI can engage with moral cognition. I have argued that based on their performance, they are able to engage with moral cognition to some extent. However, the way they represent moral properties is fundamentally different from human moral learning.

What, then, is the implication of my thesis? If emotion is not the whole story for moral cognition, we do not have a domain-specific mechanism for moral cognition, evolution does not provide us with a moral faculty, and our general rational learning mechanism such as Bayesian inferencing also fails to provide a satisfactory model for moral cognition, what is left? In my thesis, I do not intend to provide a positive account of moral cognition. What I am trying to do here is to argue that it is time to give up on a unified account for moral cognition, and also that we are not Bayesian learners when it comes to moral learning. Hence, future research on moral representation should look for the potentially many varied learning mechanisms that facilitate moral cognition.

# References

Aharoni, Eyal, Walter Sinnott-Armstrong, and Kent A. Kiehl. 2012. “Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction.” *Journal of Abnormal Psychology* 121 (2): 484–97. https://doi.org/10.1037/a0024796.

———. 2014. “What’s Wrong? Moral Understanding in Psychopathic Offenders.” *Journal of Research in Personality* 53: 175–81. https://doi.org/10.1016/j.jrp.2014.10.002.

Amiez, Céline, Charles Verstraete, Jérôme Sallet, Fadila Hadj-Bouziane, Suliann Ben Hamed, Adrien Meguerditchian, Emmanuel Procyk, et al. 2023. “The Relevance of the Unique Anatomy of the Human Prefrontal Operculum to the Emergence of Speech.” *Communications Biology* 6 (1): 693. https://doi.org/10.1038/s42003-023-05066-9.

Anderson, Rajen A., Molly J. Crockett, and David A. Pizarro. 2020. “A Theory of Moral Praise.” *Trends in Cognitive Sciences*, 1–10. https://doi.org/10.1016/j.tics.2020.06.008.

Ayer, Alfred Jules. 1936. *Language, Truth and Logic*. Vol. 47. V. Gollancz.

Baillargeon, Renée, Rose M. Scott, and Zijing He. 2010. “False-Belief Understanding in Infants.” *Trends in Cognitive Sciences* 14 (3): 110–18. https://doi.org/10.1016/j.tics.2009.12.006.

Baron-Cohen, Simon. 2019. “Autism and Mindblindness.” In *Mindblindness*. https://doi.org/10.7551/mitpress/4635.003.0010.

Baron-Cohen, Simon, Alan M. Leslie, and Uta Frith. 1985. “Does the Autistic Child Have a ‘Theory of Mind’ ?” *Cognition* 21 (1): 37–46. https://doi.org/10.1016/0010-0277(85)90022-8.

Baumard, Nicolas, Jean Baptiste André, and Dan Sperber. 2013. “A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice.” *Behavioral and Brain Sciences* 36 (1): 59–78. https://doi.org/10.1017/S0140525X11002202.

Bender, Emily M, and Alexander Koller. 2020. “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data,” 5185–98. https://www.nytimes.com/2018/11/18/technology/artific.

Bestel, Sheahan, Yingjian Bao, Hua Zhong, Xingcan Chen, and Li Liu. 2018. “Wild Plant Use and Multi-Cropping at the Early Neolithic Zhuzhai Site in the Middle Yellow River Region, China.” *Holocene* 28 (2): 195–207. https://doi.org/10.1177/0959683617721328/ASSET/IMAGES/LARGE/10.1177\_0959683617721328-FIG2.JPEG.

Białek, Michał, Rafał Muda, Jonathan Fugelsang, and Ori Friedman. 2021. “Disgust and Moral Judgment: Distinguishing Between Elicitors and Feelings Matters.” *Social Psychological and Personality Science* 12 (3): 304–13. https://doi.org/10.1177/1948550620919569.

Bian, Lin, Stephanie Sloane, and Renée Baillargeon. 2018. “Infants Expect Ingroup Support to Override Fairness When Resources Are Limited” 115 (11): 2705–10. https://www.pnas.org/content/115/11/2705.

Birch, Jonathan. 2021. “Toolmaking and the Evolution of Normative Cognition.” *Biology and Philosophy* 36 (1): 1–26. https://doi.org/10.1007/S10539-020-09777-9/FIGURES/1.

Blair, James. 1995. “A Cognitive Developmental Approach to Morality: Investigating the Psychopath.” *Cognition* 57 (1): 1–29. https://doi.org/10.1016/0010-0277(95)00676-P.

———. 1996. “Brief Report: Morality in the Autistic Child.” *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/BF02172277.

———. 1997. “Moral Reasoning and the Child with Psychopathic Tendencies.” *Personality and Individual Differences*. https://doi.org/10.1016/S0191-8869(96)00249-8.

———. 2017. “Emotion-Based Learning Systems and the Development of Morality.” *Cognition* 167: 38–45. https://doi.org/10.1016/j.cognition.2017.03.013.

Blair, James, and R. J.R. Blair. 1995. “A Cognitive Developmental Approach to Morality: Investigating the Psychopath.” *Cognition* 57 (1): 1–29. https://doi.org/10.1016/0010-0277(95)00676-P.

Blair, James, Lawrence Jones, Fiona Clark, and Margaret Smith. 1997. “The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?” *Psychophysiology* 34 (2): 192–98. https://doi.org/10.1111/j.1469-8986.1997.tb02131.x.

Bloom, Paul, and Tim P German. 2000. “Two Reasons to Abandon the False Belief Task as a Test of Theory of Mind.” *Cognition* 77 (1): B25–31. https://doi.org/https://doi.org/10.1016/S0010-0277(00)00096-2.

Boë, Louis-Jean, Thomas R Sawallis, Joël Fagot, Pierre Badin, Guillaume Barbier, Guillaume Captier, Lucie Ménard, Jean-Louis Heim, and Jean-Luc Schwartz. 2023. “Which Way to the Dawn of Speech?: Reanalyzing Half a Century of Debates and Data in Light of Speech Science.” *Science Advances* 5 (12): eaaw3916. https://doi.org/10.1126/sciadv.aaw3916.

Boyd, Robert. 2019. “A Different Kind of Animal: How Culture Transformed Our Species.” Princeton University Press. https://doi.org/10.23943/princeton/9780691195902.001.0001.

Boyd, Robert, and Peter J Richerson. 1985. *Culture and the Evolutionary Process.* *Culture and the Evolutionary Process.* Chicago, IL, US: University of Chicago Press.

Broca, Paul. 1861. “Remarques Sur Le Siège de La Faculté Du Langage Articulé, Suivies d’une Observation d’aphémie (Perte de La Parole).” *Bulletin et Memoires de La Societe Anatomique de Paris* 6: 330–57.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4,” March. http://arxiv.org/abs/2303.12712.

Cameron, C. Daryl, Lasana T. Harris, and B. Keith Payne. 2016. “The Emotional Cost of Humanity.” *Social Psychological and Personality Science* 7 (2): 105–12. https://doi.org/10.1177/1948550615604453.

Cameron, C. Daryl, B. Keith Payne, and John M. Doris. 2013. “Morality in High Definition: Emotion Differentiation Calibrates the Influence of Incidental Disgust on Moral Judgments.” *Journal of Experimental Social Psychology* 49 (4): 719–25. https://doi.org/10.1016/J.JESP.2013.02.014.

Cameron, C. Daryl, B. Keith Payne, and Joshua Knobe. 2010. “Do Theories of Implicit Race Bias Change Moral Judgments?” *Social Justice Research 2010 23:4* 23 (4): 272–89. https://doi.org/10.1007/S11211-010-0118-Z.

Carruthers, Peter. 2006. “The Architecture of the Mind.” Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199207077.001.0001.

Cesana-Arlotti, Nicolò, Ágnes Melinda Kovács, and Ernő Téglás. 2020. “Infants Recruit Logic to Learn about the Social World.” *Nature Communications 2020 11:1* 11 (1): 1–9. https://doi.org/10.1038/s41467-020-19734-5.

Chalik, Lisa, and Marjorie Rhodes. 2020. “Chapter Three - Groups as Moral Boundaries: A Developmental Perspective.” In , edited by Janette B B T - Advances in Child Development and Behavior Benson, 58:63–93. JAI. https://doi.org/https://doi.org/10.1016/bs.acdb.2020.01.003.

Chomsky, Noam. 1959. “Review of Verbal Behavior.” Edited by B F Skinner. *Language* 35 (1): 26–58. https://doi.org/10.2307/411334.

———. 1976. *Reflections on Language*. Temple Smith London.

———. 1980. *Rules and Representations*. Columbia University Press.

———. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Greenwood Publishing Group.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. “Deep Reinforcement Learning from Human Preferences.” In *Advances in Neural Information Processing Systems*, edited by I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper\_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Churchland, Patricia Smith. 1987. “Epistemology in the Age of Neuroscience.” *The Journal of Philosophy* 84 (10): 544–53. http://www.jstor.org/stable/2026917.

Churchland, Patricia Smith, and Rick Grush. 1999. “Computation and the Brain.” *The MIT Encyclopedia of Cognitive Sciences*, 155–58.

Churchland, Paul M. 1988a. “Reductionism, Connectionism, and the Plasticity of Human Consciousness.” *Cultural Dynamics* 1 (1): 29–45.

———. 1988b. “Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor.” *Philosophy of Science* 55 (2): 167–87. http://www.jstor.org/stable/187956.

Cima, Maaike, Franca Tonnaer, and Marc D. Hauser. 2010. “Psychopaths Know Right from Wrong but Don’t Care.” *Social Cognitive and Affective Neuroscience* 5 (1): 59–67. https://doi.org/10.1093/scan/nsp051.

Cosmides, Leda, Ricardo Andrés Guzmán, and John Tooby. 2018. “The Evolution of Moral Cognition.” *The Routledge Handbook of Moral Epistemology*, January, 174–228. https://doi.org/10.4324/9781315719696-10/EVOLUTION-MORAL-COGNITION-LEDA-COSMIDES-RICARDO-ANDR.

Cowie, Fiona. 1999. *What’s within?: Nativism Reconsidered*. Oxford University Press.

Crawford, Jarret T., Yoel Inbar, and Victoria Maloney. 2014. “Disgust Sensitivity Selectively Predicts Attitudes toward Groups That Threaten (or Uphold) Traditional Sexual Morality.” *Personality and Individual Differences* 70: 218–23. https://doi.org/10.1016/J.PAID.2014.07.001.

Crockett, Molly J. 2013. “Models of Morality.” *Trends in Cognitive Sciences*. Elsevier. https://doi.org/10.1016/j.tics.2013.06.005.

Csikszentmihalyi, Mark. 2020. “Confucius.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Summer 202. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/confucius/.

Curry, Oliver Scott. 2016. “Morality as Cooperation: A Problem-Centred Approach,” 27–51. https://doi.org/10.1007/978-3-319-19671-8\_2.

Curry, Oliver Scott, Mark Alfano, Mark J. Brandt, and Christine Pelican. 2021. “Moral Molecules: Morality as a Combinatorial System.” *Review of Philosophy and Psychology*, August, 1–20. https://doi.org/10.1007/S13164-021-00540-X/TABLES/2.

Cushman, Fiery. 2013. “Action, Outcome, and Value: A Dual-System Framework for Morality.” *Personality and Social Psychology Review* 17 (3): 273–92. https://doi.org/10.1177/1088868313495594.

D’Arms, Justin. 2014. “Sentimentalism and Scientism.” In *Moral Psychology and Human Agency*. https://doi.org/10.1093/acprof:oso/9780198717812.003.0011.

D’Arms, Justin, and Daniel Jacobson. 2000. “The Moralistic Fallacy: On the ‘Appropriateness’ of Emotions.” *Philosophy and Phenomenological Research* 61 (1): 65. https://doi.org/10.2307/2653403.

Demaree-Cotton, Joanna, and Guy Kahane. 2018. “The Neuroscience of Moral Judgment.” In *The Routledge Handbook of Moral Epistemology*. https://doi.org/10.4324/9781315719696-5.

Dewey, John. 1895. “The Theory of Emotion.” *Psychological Review* 2 (1).

Downes, Stephen M. 2021. “Evolutionary Psychology.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Spring 202. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/evolutionary-psych ology/.

Dwyer, Susan. 1999. “Reconciliation for Realists.” *Ethics & International Affairs* 13: 81–98.

———. 2006. “How Good Is the Linguistic Analogy?” In *The Innate Mind: Structure and Contents*, edited by Peter Carruthers, Stephen Laurence, and Stephen P Stich. Oxford University Press.

Dwyer, Susan, Bryce Huebner, and Marc D Hauser. 2010. “The Linguistic Analogy: Motivations, Results, and Speculations.” *Topics in Cognitive Science* 2 (3): 486–510.

Earp, Brian D., Killian L. McLoughlin, Joshua T. Monrad, Margaret S. Clark, and Molly J. Crockett. 2021. “How Social Relationships Shape Moral Wrongness Judgments.” *Nature Communications 2021 12:1* 12 (1): 1–13. https://doi.org/10.1038/s41467-021-26067-4.

Engelmann, Jan M., and Michael Tomasello. 2019. “Children’s Sense of Fairness as Equal Respect.” *Trends in Cognitive Sciences* 23 (6): 454–63. https://doi.org/10.1016/J.TICS.2019.03.001.

Eskine, Kendall J., Natalie A. Kacinik, and Jesse J. Prinz. 2011. “A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment.” *Psychological Science* 22 (3): 295–99. https://doi.org/10.1177/0956797611398497.

Everett, Daniel L. 2005. “Cultural Constraints on Grammar and Cognition in Pirahã Another Look at the Design Features of Human Language.” *Current Anthropology* 46 (4): 621–46. https://doi.org/10.1086/431525.

Fedorenko, Evelina, and Idan A Blank. 2020. “Broca’s Area Is Not a Natural Kind.” *Trends in Cognitive Sciences* 24 (4): 270–84.

Fedorenko, Evelina, and Rosemary Varley. 2016. “Language and Thought Are Not the Same Thing: Evidence from Neuroimaging and Neurological Patients.” *Annals of the New York Academy of Sciences* 1369 (1): 132–53. https://doi.org/https://doi.org/10.1111/nyas.13046.

FeldmanHall, Oriel, Jae-Young Son, and Joseph Heffner. 2018. “Norms and the Flexibility of Moral Action.” *Personality Neuroscience* 1 (September). https://doi.org/10.1017/PEN.2018.13.

Fischer, John Martin. 1999. “Recent Work on Moral Responsibility.” *Ethics* 110 (1): 93–139.

Fitouchi, Léo, Jean-Baptiste André, and Nicolas Baumard. 2022. “Moral Disciplining: The Cognitive and Evolutionary Foundations of Puritanical Morality.” *Behavioral and Brain Sciences*, September, 1–71. https://doi.org/10.1017/S0140525X22002047.

Floridi, Luciano, and J.W. Sanders. 2004. “On the Morality of Artificial Agents.” *Minds and Machines* 14 (3): 349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Fodor, Jerry A. 1983. *The Modularity of Mind*. MIT press.

Foot, Philippa. 1967. “The Problem of Abortion and the Doctrine of the Double Effect.” *Oxford Review* 5.

Frankfurt, Harry. 1969. “Alternate Possibilities and Moral Responsibility.” *Journal of Philosophy* 66 (23).

Gert, Bernard, and Joshua Gert. 2017. “The Definition of Morality.” In *The Stanford Encyclopedia of Philosophy*, Fall 2017. Metaphysics Research Lab, Stanford University. https://doi.org/10.5860/choice.41sup-0181.

Ghelfi, Eric, Cody D. Christopherson, Heather L. Urry, Richie L. Lenne, Nicole Legate, Mary Ann Fischer, Fieke M. A. Wagemans, et al. 2020. “Reexamining the Effect of Gustatory Disgust on Moral Judgment: A Multilab Direct Replication of Eskine, Kacinik, and Prinz (2011):” *Https://Doi.Org/10.1177/2515245919881152* 3 (1): 3–23. https://doi.org/10.1177/2515245919881152.

Gilbert, Daniel T, Elizabeth C Pinel, Timothy D Wilson, Stephen J Blumberg, and Thalia P Wheatley. 1998. “Immune Neglect: A Source of Durability Bias in Affective Forecasting.” *Journal of Personality and Social Psychology*. US: American Psychological Association. https://doi.org/10.1037/0022-3514.75.3.617.

Goodwin, Geoffrey P, and John M Darley. 2008. “The Psychology of Meta-Ethics: Exploring Objectivism.” *Cognition* 106 (3).

Green, Mitchell S. 2010. “Speech Acts.” In *Stanford Encyclopedia of Philosophy*, edited by Mitchell Green.

Greene, Joshua. 2014. “Moral Tribes.” *Emotion, Reason and the Gap Between Us and Them*. https://doi.org/10.1017/CBO9781107415324.004.

Greene, Joshua, Leigh Nystrom, Andrew Engell, John Darley, and Jonathan Cohen. 2004. “The Neural Bases of Cognitive Conflict and Control in Moral Judgment.” *Neuron* 44 (2): 389–400. https://doi.org/https://doi.org/10.1016/j.neuron.2004.09.027.

Greene, Joshua, Brian Sommerville, Leigh Nystrom, John Darley, and Jonathan Cohen. 2001. “An FMRI Investigation of Emotional Engagement in Moral Judgment.” *Science* 293 (5537): 2105 LP – 2108. https://doi.org/10.1126/science.1062872.

Greene, Joshuad., Brian Sommerville, Leigh E. Nystrom, John M. Darley, Jonathan D. Cohen, R. B. Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. “An FMRI Investigation of Emotional Engagement in Moral Judgment.” *Science* 293 (5537): 2105 LP – 2108. https://doi.org/10.1126/science.1062872.

Gweon, Hyowon. 2021. “Inferential Social Learning: Cognitive Foundations of Human Social Learning and Teaching.” *Trends in Cognitive Sciences*, August. https://doi.org/10.1016/J.TICS.2021.07.008.

Hagoort, Peter. 2005. “On Broca, Brain, and Binding: A New Framework.” *Trends in Cognitive Sciences* 9 (9): 416–23.

Haidt, Jonathan. 2001. “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review* 108 (4): 814–34. https://doi.org/10.1037/0033-295X.108.4.814.

———. 2003. “The Moral Emotions.” *Handbook of Affective Sciences* 11 (2003): 852–70.

Haidt, Jonathan, and Jesse Graham. 2007. “When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize.” *Social Justice Research* 20 (1): 98–116. https://doi.org/10.1007/S11211-007-0034-Z.

Haidt, Jonathan, Silvia Helena Koller, and Maria G. Dias. 1993a. “Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?” *Journal of Personality and Social Psychology* 65 (4): 613–28. https://doi.org/10.1037/0022-3514.65.4.613.

———. 1993b. “Is It Wrong to Eat Your Dog?” *Journal of Personality and Social Psychology* 65 (4): 613–28.

Haidt, Jonathan, Clark McCauley, and Paul Rozin. 1994. “Individual Differences in Sensitivity to Disgust: A Scale Sampling Seven Domains of Disgust Elicitors.” *Personality and Individual Differences* 16 (5): 701–13. https://doi.org/10.1016/0191-8869(94)90212-7.

Hamlin, J. Kiley, Karen Wynn, and Paul Bloom. 2007. “Social Evaluation by Preverbal Infants.” *Nature* 450 (7169): 557–59. https://doi.org/10.1038/nature06288.

Hare, R M. 1970. “Meaning and Speech Acts.” *The Philosophical Review* 79 (1): 3–24. https://doi.org/10.2307/2184066.

Hare, Robert D. 1991. *The Hare Psychopathy Checklist-Revised: Manual*. Multi-Health Systems, Incorporated.

Hauser, Marc. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong.* Ecco/HarperCollins Publishers.

Henrich, Joseph. 2016. “The Secret of Our Success : How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.”

Henrich, Joseph, Damián E. Blasi, Cameron M. Curtin, Helen Elizabeth Davis, Ze Hong, Daniel Kelly, and Ivan Kroupin. 2022. “A Cultural Species and Its Cognitive Phenotypes: Implications for Philosophy.” *Review of Philosophy and Psychology 2022*, February, 1–38. https://doi.org/10.1007/S13164-021-00612-Y.

Henrich, Joseph, and James Broesch. 2011. “On the Nature of Cultural Transmission Networks: Evidence from Fijian Villages for Adaptive Learning Biases.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1567): 1139–48. https://doi.org/10.1098/RSTB.2010.0323.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *Behavioral and Brain Sciences* 33 (2–3): 61–83. https://doi.org/10.1017/S0140525X0999152X.

Heyes, Cecilia. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press.

———. 2021. “Is Morality a Gadget? Nature, Nurture and Culture in Moral Development.” *Synthese* 198 (5): 4391–4414. https://doi.org/10.1007/S11229-019-02348-W/TABLES/1.

Hofmann, Wilhelm, Daniel C Wisneski, Mark J Brandt, and Linda J Skitka. 2014. “Morality in Everyday Life.” *Science* 345 (6202): 1340–43. https://doi.org/10.1126/science.1251560.

Hume, David. 1751. *An Enquiry Concerning the Principles of Morals*. London: London Printed for A. Millar, 1751.

Inbar, Yoel, and David Pizarro. 2021. “How Disgust Affects Social Judgments.” PsyArXiv. https://doi.org/10.31234/OSF.IO/AXHVN.

Inbar, Yoel, David A. Pizarro, and Paul Bloom. 2009. “Conservatives Are More Easily Disgusted than Liberals.” *Cognition and Emotion* 23 (4): 714–25. https://doi.org/10.1080/02699930802110007.

Ingram, David. 1989. *First Language Acquisition: Method, Description and Explanation*. Cambridge university press.

Izard, Carroll E. 2010. “More Meanings and More Questions for the Term ‘Emotion.’” *Emotion Review* 2 (4): 383–85. https://doi.org/10.1177/1754073910374670.

Jackendoff, Ray. 1999. “The Representational Structures of the Language Faculty and Their Interactions.” *The Neurocognition of Language*, 37–79.

James, William. 1884. “What Is an Emotion?” *Mind* 9 (n/a).

———. 1891. “The Moral Philosopher and the Moral Life.” *The International Journal of Ethics* 1 (3): 330–54.

Jin, Kyong sun, and Renée Baillargeon. 2017. “Infants Possess an Abstract Expectation of Ingroup Support.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (31): 8199–8204. https://doi.org/10.1073/PNAS.1706286114/-/DCSUPPLEMENTAL.

Johnson, David J., Jessica Wortman, Felix Cheung, Megan Hein, Richard E. Lucas, M. Brent Donnellan, Charles R. Ebersole, and Rachel K. Narr. 2016. “The Effects of Disgust on Moral Judgments: Testing Moderators.” *Http://Dx.Doi.Org/10.1177/1948550616654211* 7 (7): 640–47. https://doi.org/10.1177/1948550616654211.

Jylkkä, Jussi, Johanna Härkönen, and Jukka Hyönä. 2020. “Incidental Disgust Does Not Cause Moral Condemnation of Neutral Actions.” *Https://Doi.Org/10.1080/02699931.2020.1810639* 35 (1): 96–109. https://doi.org/10.1080/02699931.2020.1810639.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. macmillan.

Kant, Immanuel. 1785. *Groundwork of The Metaphysics of Morals*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511813306.007.

Kauppinen, Antti. 2019. “What Is Sentimentalism? What Is Rationalism?” *Behavioral and Brain Sciences* 42. https://doi.org/10.1017/s0140525x18002649.

Kelly, Daniel. 2011. *Yuck!: The Nature and Moral Significance of Disgust*. MIT press.

Kelly, Daniel, and Stephen Setman. 2021. “The Psychology of Normative Cognition.” In *The Stanford Encyclopedia of Philosophy*.

Kelly, Daniel, Stephen Stich, Kevin J Haley, Serena J Eng, and Daniel M T Fessler. 2007. “Harm , Affect , and the Moral / Conventional Distinction” 22 (2): 117–31.

Kiehl, Kent A. 2006. “A Cognitive Neuroscience Perspective on Psychopathy: Evidence for Paralimbic System Dysfunction.” *Psychiatry Research* 142 (2–3): 107–28. https://doi.org/10.1016/j.psychres.2005.09.013.

Kiley Hamlin, J., Tomer Ullman, Josh Tenenbaum, Noah Goodman, and Chris Baker. 2013. “The Mentalistic Basis of Core Social Cognition: Experiments in Preverbal Infants and a Computational Model.” *Developmental Science* 16 (2): 209–26. https://doi.org/10.1111/desc.12017.

Knobe, Joshua. 2003a. “Intentional Action in Folk Psychology: An Experimental Investigation.” *Philosophical Psychology* 16 (2): 309–24. https://doi.org/10.1080/09515080307771.

———. 2003b. “Intentional Action and Side Effects in Ordinary Language.” *Analysis* 63 (3): 190–94. http://www.jstor.org/stable/3329308.

Kohlberg, Lawrence. 1958. *The Development of Modes of Moral Thinking and Choice in the Years 10 to 16*. University of Chicago.

———. 1986. “A Current Statement on Some Theoretical Issues.” *Lawrence Kohlberg: Consensus and Controversy* 1: 485–546.

Kozuch, Benjamin, and Shaun Nichols. 2011. “Awareness of Unawareness.” *Journal of Consciousness Studies* 18 (11–12): 135–60.

Kumar, Victor, and Richmond Campbell. 2022. “A Better Ape: The Evolution of the Moral Mind and How It Made Us Human.” Oxford University Press. https://doi.org/10.1093/oso/9780197600122.001.0001.

Landy, Justin F., and Geoffrey P. Goodwin. 2015. “Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence.” *Perspectives on Psychological Science* 10 (4): 518–36. https://doi.org/10.1177/1745691615583128.

Lapka, Samantha P, Franki Y H Kung, Justin P Brienza, and Abigail A Scholer. 2022. “Determined Yet Dehumanized: People Higher in Self-Control Are Seen as More Robotic.” *Social Psychological and Personality Science* 14 (2): 117–29. https://doi.org/10.1177/19485506221093109.

Laurence, Stephen, and Eric Margolis. 2001. “The Poverty of the Stimulus Argument.” *British Journal for the Philosophy of Science* 52 (2).

Lazarus, Richard S. 1991. *Emotion and Adaptation [Electronic Resource]*. New York : New York .

Lenman, James. 2007. “Expressivism and Epistemology: What Is Moral Inquiry?” *Aristotelian Society Supplementary Volume* 81 (1): 63–81. https://doi.org/10.1111/J.1467-8349.2007.00151.X.

Lerner, Jennifer S, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. 2015. “Emotion and Decision Making.” *Annual Review of Psychology* 66 (1): 799–823. https://doi.org/10.1146/annurev-psych-010213-115043.

Lewens, Tim, and Andrew Buskell. 2023. “Cultural Evolution.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta and Uri Nodelman, Summer 202. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/evolution-cultural /.

Liu, Li. 2021. “Communal Drinking Rituals and Social Formations in the Yellow River Valley of Neolithic China.” *Journal of Anthropological Archaeology* 63 (September): 101310. https://doi.org/10.1016/j.jaa.2021.101310.

Maclure, Jocelyn. 2021. “AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind.” *Minds and Machines 2021 31:3* 31 (3): 421–38. https://doi.org/10.1007/S11023-021-09570-X.

Malle, Bertram F. 2021. “Moral Judgments.” *Annual Review of Psychology* 72 (1): 293–318. https://doi.org/10.1146/annurev-psych-072220-104358.

Mandelbaum, Eric. 2019. “Troubles with Bayesianism: An Introduction to the Psychological Immune System.” *Mind & Language* 34 (2): 141–57. https://doi.org/10.1111/MILA.12205.

Margolis, Eric, and Stephen Laurence. 2022. “Concepts.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta and Uri Nodelman, Fall 2022. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/concepts/.

Marr, David. 1982. “Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, Henry Holt and Co.” *Inc., New York, NY* 2 (4.2).

Marr, David, and Herbert Keith Nishihara. 1978. “Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes.” *Proceedings of the Royal Society of London. Series B. Biological Sciences* 200 (1140): 269–94.

Maschio, Nicola Del, Davide Fedeli, Gioacchino Garofalo, and Giovanni Buccino. 2022. “Evidence for the Concreteness of Abstract Language: A Meta-Analysis of Neuroimaging Studies.” *Brain Sciences*. https://doi.org/10.3390/brainsci12010032.

Matúš, Pikuliak. 2023. “ChatGPT Survey: Performance on NLP Datasets.” 2023. https://www.opensamizdat.com/posts/chatgpt\_survey.

May, Joshua. 2014. “Does Disgust Influence Moral Judgment?” *Australasian Journal of Philosophy* 92 (1): 125–41. https://doi.org/10.1080/00048402.2013.797476.

———. 2018. *Regard for Reason in the Moral Mind*. *Regard for Reason in the Moral Mind*. https://doi.org/10.1093/oso/9780198811572.001.0001.

May, Joshua, and Victor Kumar. 2018. “Moral Reasoning and Emotion.” *The Routledge Handbook of Moral Epistemology*.

McHoskey, John W. 1995. “Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization.” *Basic and Applied Social Psychology* 17 (3): 395–409. https://doi.org/10.1207/s15324834basp1703\_7.

Mikhail, John. 2007. “Universal Moral Grammar : Theory , Evidence and the Future” 11 (4). https://doi.org/10.1016/j.tics.2006.12.007.

———. 2011. *Elements of Moral Cognition: Rawls’ Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press.

Moon, Jordan W, Jaimie Arona Krems, and Adam B Cohen. 2018. “Religious People Are Trusted Because They Are Viewed as Slow Life-History Strategists.” *Psychological Science* 29 (6): 947–60.

Nelson, Sharon A. 1980. “Factors Influencing Young Children’s Use of Motives and Outcomes as Moral Criteria.” *Child Development*, 823–29.

Nettle, Daniel, and Rebecca Saxe. 2020. “Preferences for Redistribution Are Sensitive to Perceived Luck, Social Homogeneity, War and Scarcity.” *Cognition* 198: 104234.

Neumann, John Von, and Oskar Morgenstern. 1947. “Theory of Games and Economic Behavior, 2nd Rev.”

Neumann, Sigmund. 1949. “The International Civil War.” *World Politics* 1 (3): 333–50. https://doi.org/DOI: 10.2307/2009034.

Nichols, Shaun. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgement*. *Sentimental Rules: On the Natural Foundations of Moral Judgement*. https://doi.org/10.1093/0195169344.001.0001.

———. 2021. *Rational Rules: Towards a Theory of Moral Learning*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198869153.001.0001.

Nichols, Shaun, and Joshua Knobe. 2007. “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions.” *Nous*. https://doi.org/10.1111/j.1468-0068.2007.00666.x.

Nisbett, Richard E., and Timothy D. Wilson. 1977a. “Telling More than We Can Know: Verbal Reports on Mental Processes.” *Psychological Review*. https://doi.org/10.1037/0033-295X.84.3.231.

———. 1977b. “The Halo Effect: Evidence for Unconscious Alteration of Judgments.” *Journal of Personality and Social Psychology*. https://doi.org/10.1037//0022-3514.35.4.250.

Nussbaum, Martha. 2001. *Upheavals of Thought : The Intelligence of Emotions*. Cambridge: Cambridge.

Olson, Jonas. 2010. “In Defense of Moral Error Theory.” In *New Waves in Metaethics*, edited by Michael Brady. Palgrave-Macmillan.

OpenAI. 2023. “GPT-4 Technical Report,” March. http://arxiv.org/abs/2303.08774.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. “Training Language Models to Follow Instructions with Human Feedback,” March. http://arxiv.org/abs/2203.02155.

Parkinson, Carolyn, Walter Sinnott-armstrong, Philipp.E Koralus, Angela Mendelovici, Victoria Mcgeer, and Thalia Wheatley. 2011. “Is Morality Unified? Evidence That Distinct Neural Systems Underlie Moral Judgments of Harm, Dishonesty, and Disgust,” 3162–80. https://doi.org/10.1162/jocn\_a\_00017.

Partington, Scott, Shaun Nichols, and Tamar Kushnir. 2023. “Rational Learners and Parochial Norms.” *Cognition* 233 (April): 105366. https://doi.org/10.1016/J.COGNITION.2022.105366.

Petkov, Christopher, and Erich Jarvis. 2012. “Birds, Primates, and Spoken Language Origins: Behavioral Phenotypes and Neurobiological Substrates .” *Frontiers in Evolutionary Neuroscience* .

Pinker, Steven, and Paul Bloom. 1990. “Natural Language and Natural Selection.” *Behavioral and Brain Sciences* 13 (4): 707–27. https://doi.org/10.1017/S0140525X00081061.

Pinker, Steven, and Jerry Fodor. 2005. “So How Does the Mind Work?” *Mind & Language* 20 (1): 1–24. https://doi.org/10.1111/J.0268-1064.2005.00274.X.

Porot, Nicolas, and Eric Mandelbaum. 2021. “The Science of Belief: A Progress Report.” *Wiley Interdisciplinary Reviews: Cognitive Science* 12 (2): e1539. https://doi.org/10.1002/WCS.1539.

Prinz, Jesse. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. *Oxford University Press*. https://doi.org/10.1093/mind/fzl450.

———. 2006. “The Emotional Basis of Moral Judgments.” *Philosophical Explorations* 9 (1): 29–43. https://doi.org/10.1080/13869790500492466.

———. 2007. *The Emotional Construction of Morals*. Oxford University Press.

———. 2012. “The Moral Emotions.” In *The Oxford Handbook of Philosophy of Emotion*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199235018.003.0024.

———. 2016. “Sentimentalism and the Moral Brain.” In *Moral Brains*, 45–73. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199357666.003.0002.

Prinz, Jesse, and Shaun Nichols. 2010. “Moral Emotions.” In *The Moral Psychology Handbook*. Oxford University Press.

Quintelier, Katinka. 2010. “Moral Relativism and the Moral / Conventional Distinction,” 1–11.

Railton, Peter. 2014. “The Affective Dog and Its Rational Tale: Intuition and Attunement.” *Ethics* 124 (4): 813–59. https://doi.org/10.1086/675876.

Rawls, John. 1971. “A Theory of Justice (Original Edition).” *Development Policy Review*.

———. 2020. *A Theory of Justice: Revised Edition*. Harvard university press.

Rhodes, Marjorie, and Kelsey Moty. 2020. “What Is Social Essentialism and How Does It Develop?” In *Advances in Child Development and Behavior*, 59:1–30. Academic Press Inc. https://doi.org/10.1016/bs.acdb.2020.05.001.

Robbins, Philip. 2017. “Modularity of Mind.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta, Winter 201. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/modularity-mind/.

Roese, Neal J, and James M Olson. 1997. “Counterfactual Thinking: The Intersection of Affect and Function.” In , edited by Mark P B T - Advances in Experimental Social Psychology Zanna, 29:1–59. Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2601(08)60015-5.

Rosati, Connie S. 2016. “Moral Motivation.” *Stanford Encyclopedia of Philosophy*.

Röseler, Lukas, Jacqueline Ebert, Astrid Schütz, and Roy Baumeister. 2021. “The Upsides and Downsides of High Self-Control: Evidence for Effects of Similarity and Situation Dependency.” *Europe’s Journal of Psychology* 17 (1 SE-Research Reports): 1–16. https://doi.org/10.5964/ejop.2639.

Ross, Don. 2005. *Economic Theory and Cognitive Science: Microexplanation*. MIT press.

Royzman, Edward B, Geoffrey P Goodwin, and Robert F Leeman. 2011. “When Sentimental Rules Collide: ‘Norms with Feelings’ in the Dilemmatic Context.” *Cognition* 121 (1): 101–14. https://doi.org/https://doi.org/10.1016/j.cognition.2011.06.006.

Rozin, Paul, Jonathan Haidt, and Clark R McCauley. 2008. “Disgust.” In *Handbook of Emotions, 3rd Ed.*, 757–76. New York, NY, US: The Guilford Press.

Russell, Stuart, and Peter Norvig. 2021. *Artificial Intelligence : A Modern Approach*. 4th ed.

Sanyal, Mohima, William H. B. McAuliffe, and Oliver Scott Curry. 2021. “Gross Values: Investigating the Role of Disgust in Bioethics.” *Current Psychology 2021*, March, 1–8. https://doi.org/10.1007/S12144-021-01609-7.

Schlosser, Markus. 2019. “Agency.” In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=agency.

Schnall, Simone, Jennifer Benton, and Sophie Harvey. 2008. “With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments.” *Psychological Science*. https://doi.org/10.1111/j.1467-9280.2008.02227.x.

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2008. “Disgust as Embodied Moral Judgment.” *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167208317771.

———. 2015. “Landy and Goodwin (2015) Confirmed Most of Our Findings Then Drew the Wrong Conclusions.” *Perspectives on Psychological Science* 10 (4): 537–38. https://doi.org/10.1177/1745691615589078.

Schroeder, Timothy, Adina L Roskies, and Shaun Nichols. 2010. “Moral Motivation.” In *Moral Psychology Handbook*, edited by John Doris. Oxford University Press.

Schroeter, François, and Karen Jones. 2018. “The Many Moral Rationalisms.” *The Many Moral Rationalisms*, no. November 2019: 1–309. https://doi.org/10.1093/oso/9780198797074.001.0001.

Schubbach, Arno. 2019. “Judging Machines: Philosophical Aspects of Deep Learning.” *Synthese 2019 198:2* 198 (2): 1807–27. https://doi.org/10.1007/S11229-019-02167-Z.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. “Proximal Policy Optimization Algorithms,” July. http://arxiv.org/abs/1707.06347.

Sell, Aaron, Daniel Sznycer, Laith Al-Shawaf, Julian Lim, Andre Krauss, Aneta Feldman, Ruxandra Rascanu, Lawrence Sugiyama, Leda Cosmides, and John Tooby. 2017. “The Grammar of Anger: Mapping the Computational Architecture of a Recalibrational Emotion.” *Cognition* 168 (November): 110–28. https://doi.org/10.1016/j.cognition.2017.06.002.

Sinnott-Armstrong, Walter. 2006. *Moral Skepticism*. Oxford University Press.

Skinner, B F. 1957. *Verbal Behavior*. Appleton-Century-Crofts.

Slote, Michael. 2010. *Moral Sentimentalism*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195391442.001.0001.

Smith, Michael. 1994. *The Moral Problem*. Blackwell.

———. 2018. “Three Kinds of Moral Rationalism.” In *The Many Moral Rationalisms*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198797074.003.0003.

Sober, Elliott. 1992. “Models of Cultural Evolution BT - Trees of Life: Essays in Philosophy of Biology.” In , edited by Paul Griffiths, 17–39. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-8038-0\_2.

Sperber, Dan, and Nicolas Baumard. 2012. “Moral Reputation: An Evolutionary and Cognitive Perspective” 27 (5): 495–518. https://doi.org/10.1111/MILA.12000.

Sripada, Chandra Sekhar. 2008. “Nativism and Moral Psychology: Three Models of the Innate Structure That Shapes the Contents of Moral Norms.” In *Moral Psychology, Volume 1: The Evolution of Morality: Adaptations and Innateness*, edited by Walter Sinnott-armstrong, 319–43. https://doi.org/doi.org/10.7551/mitpress/7481.003.0008.

Sripada, Chandra Sekhar, and Stephen Stich. 2007. “A Framework for the Psychology of Norms.” Edited by Peter Carruthers, Stephen Laurence, and Stephen P. Stich. *Innateness and the Structure of the Mind, Volume II* II: 280–301. https://philpapers.org/rec/STIAFF.

Stevenson, Charles Leslie. 1944. *Ethics and Language*. Yale University Press.

Stich, Stephen. 2006. “Is Morality an Elegant Machine or a Kludge?” *Journal of Cognition and Culture* 6 (1–2): 181–89. https://doi.org/https://doi.org/10.1163/156853706776931349.

———. 2018. “Do We Really Externalize or Objectivize Moral Demands?” *The Behavioral and Brain Sciences* 41: e113. https://doi.org/10.1017/S0140525X18000183.

Strawson, Peter. 1962. “Freedom and Resentment.” In *Proceedings of the British Academy, Volume 48: 1962*, edited by Gary Watson. Vol. 48. Oup Oxford.

Thomson, Judith Jarvis. 1976. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59 (2): 204–17.

Ting, Fransisca, and Renée Baillargeon. 2021. “Toddlers Draw Broad Negative Inferences from Wrongdoers’ Moral Violations.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (39). https://doi.org/10.1073/PNAS.2109045118/-/DCSUPPLEMENTAL.

Ting, Fransisca, Zijing He, and Renée Baillargeon. 2019. “Toddlers and Infants Expect Individuals to Refrain from Helping an Ingroup Victim’s Aggressor.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (13): 6025–34. https://doi.org/10.1073/PNAS.1817849116/-/DCSUPPLEMENTAL.

Tomasello, Michael. 2020. “The Moral Psychology of Obligation.” *Behavioral and Brain Sciences* 43. https://doi.org/10.1017/S0140525X19001742.

Tooby, John, and Leda Cosmides. 2005. “Evolutionary Psychology: Conceptual Foundations.” *The Handbook of Evolutionary Psychology*, 5–67.

Turiel, Elliot. 1983. *The Development of Social Knowledge. Morality and Convention*. Vol. 33. Cambridge University Press.

Tversky, Amos, and Daniel Kahneman. 1974. “Judgment under Uncertainty: Heuristics and Biases.” *Science* 185 (4157): 1124–31. https://doi.org/10.1126/science.185.4157.1124.

———. 1983. “Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.” *Psychological Review* 90 (4): 293–315. https://doi.org/10.1037/0033-295X.90.4.293.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems* 2017-Decem (June): 5999–6009. https://arxiv.org/abs/1706.03762v5.

Warneken, Felix, Maria Gräfenhain, and Michael Tomasello. 2012. “Collaborative Partner or Social Tool? New Evidence for Young Children’s Understanding of Joint Intentions in Collaborative Activities.” *Developmental Science* 15 (1): 54–61. https://doi.org/10.1111/J.1467-7687.2011.01107.X.

Way, Pluralist, Richard A Shweder, and Jonathan Haidt. 1993. “Commentary to Feature Review: The Future of Moral Psychology: Truth, Intuition, and The.” *Source: Psychological Science* 4 (6): 360–65.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. “Emergent Abilities of Large Language Models,” June. http://arxiv.org/abs/2206.07682.

Wheatley, Thalia, and Jonathan Haidt. 2005. “Hypnotic Disgust Makes Moral Judgments More Severe.” *Psychological Science* 16 (10): 780–84.

Xu, Fei, and Joshua B. Tenenbaum. 2007. “Word Learning as Bayesian Inference.” *Psychological Review* 114 (2): 245–72. https://doi.org/10.1037/0033-295X.114.2.245.

Zerilli, John. 2022. “Explaining Machine Learning Decisions.” *Philosophy of Science* 89 (1): 1–19. https://doi.org/10.1017/psa.2021.13.

1. Some researchers such as Jonathan Haidt use the terms ‘intuition’ and ‘emotion’ interchangeably in their argument. [↑](#footnote-ref-2)
2. Similarly, D’Arms, who claims himself as ‘rational sentimentalist’, argues that reasoning can regulate our moral emotions(D’Arms 2014, 256). [↑](#footnote-ref-3)
3. Nichols cites an extensive range of psychological experiments, however these two experiments are the most frequently cited and serve as key examples for Nichols to argue against his opponents. [↑](#footnote-ref-4)
4. As far as I understand, Nichols has identified ToM with Empathy, but I intend to talk about them separately. [↑](#footnote-ref-5)
5. This is a revised version of the original one, they are similar in general. [↑](#footnote-ref-6)
6. I will follow Marr’s understanding of representation (Marr 1982). According to Marr (1982, 21), a representation is a formal system for ‘making explicit certain entities or types of types of information, together with a specification of how the system does this’ (Marr 1982, 21). Accordingly, the result of using a representation to describe a given entity is a *description* of the entity in that representation (Marr and Nishihara 1978). In other words, representation can be understood as a functional system, in which it captures some features of reality and describe them by using a symbol. [↑](#footnote-ref-7)
7. In light of Marr’s interpretation of representation, moral representation refers to a cognitive formal system that detects moral cues from reality, and describes these moral cues by using a set of symbols, which are probably some sets of moral codes. [↑](#footnote-ref-8)
8. There is an ongoing debate in animal minds on whether non-animals have linguistic capacity(Petkov and Jarvis 2012). I will not engage in the debate here. I will just assume that humans demonstrate a unique and powerful linguistic capacity. [↑](#footnote-ref-9)
9. It is worth making a difference here: the claim that linguistic capacity is realised in Broca’s area doesn’t mean that Broca’s area only deals with linguistic information-processing. Therefore, it won’t be an sufficient objection if Broca’s area also deals with other information-processing tasks(Fedorenko and Blank 2020). [↑](#footnote-ref-10)
10. My demonstration of moral rationalism here is simplified, and does not illustrate some fundamental differences between each version of moral rationalism, because these differences are not relevant to this chapter. But it is worth noting that moral rationalists do disagree on the nature of moral reasons. According to Smith (Smith 2018), there are two ways in which we can understand the relationship between moral reasons (also moral facts, as moral facts are reasons for acting morally) and general reasons for actions. One way is to argue that moral facts and reasons for actions are two distinct domains, as moral facts are predetermined, while general reasons for actions are not responsive to empirical evidence. Another way is to argue that moral facts are a kind of sub-domain of reason for action. No matter which way, moral rationalists all presuppose that moral facts are universally true, impartial and unconditional, the difference just lies in how they will justify this view. [↑](#footnote-ref-11)
11. Since rational rules account scaffolds its theoretical framework on the basis of Bayesianism, I will also call rational rules account moral Bayesianism in this chapter. [↑](#footnote-ref-12)
12. Sripada and Stich’s framework also implies a domain-specific mechanism for morality, which I disagree with. Regardless, we can still infer that a moral cognitive system needs to achieve these functions and representation no matter it is a domain-specific or not. [↑](#footnote-ref-13)