



AMERICAN UNIVERSITY OF BEIRUT

MORAL FOLK INTUITIONS: PSYCHOLOGICAL  
FOUNDATIONS AND NORMATIVE IMPLICATIONS

by  
FATIMA HUSSEIN SADEK

A thesis  
submitted in partial fulfillment of the requirements  
for the degree of Master of Arts  
to the Department of Philosophy  
of the Faculty of Arts and Sciences  
at the American University of Beirut

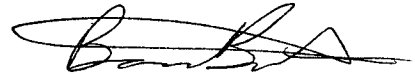
Beirut, Lebanon  
August 2019

AMERICAN UNIVERSITY OF BEIRUT

MORAL FOLK INTUITIONS: PSYCHOLOGICAL  
FOUNDATIONS AND NORMATIVE IMPLICATIONS


by  
FATIMA HUSSEIN SADEK

Approved by:



Dr. Bana Bashour, Associate Professor  
Philosophy

Advisor



Dr. Quinn Gibson, Assistant Professor  
Philosophy

Member of Committee



Dr. Bashshar H. Haydar, Professor  
Philosophy

Member of Committee

Date of thesis defense: [August 1<sup>st</sup>, 2019]

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name:

Sadek Fatima Hussein  
Middle Last First

☒ Master's Thesis

☐ Master's Project

☐ Doctoral Dissertation

☒ I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

☐ I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after:

**One ---- year from the date of submission of my thesis, dissertation, or project.**

**Two ---- years from the date of submission of my thesis, dissertation, or project.**

**Three ---- years from the date of submission of my thesis, dissertation, or project.**

Fatima 9<sup>th</sup> of August, 2019  
Signature Date

## ACKNOWLEDGMENTS

I would like to first thank my thesis advisor, Dr. Bana Bashour, for her support and hard work in making sure that I finish my work on time and do so as efficiently as possible. Thank you for your time, motivation, constant presence and valuable assistance. Without you, this thesis would not have been possible.

I would also like to thank my committee members, Dr. Bashshar Haydar and Dr. Quinn Gibson, for the encouraging and helpful comments and feedback, for being accommodating of my needs and requests and for being there every step of the way in this process. Your help and guidance was indispensable.

Lastly, I would like to thank my parents for supporting me throughout this long journey. Without their encouragement, support and care, I wouldn't have been able to keep up with this program. I would also like to thank my friends for their emotional and motivational support. Every kind word, every encouragement helped me keep going forward.

# AN ABSTRACT OF THE THESIS OF

Fatima Hussein Sadek for Master of Art  
Major: Philosophy

Title: Moral Folk Intuitions: Psychological Foundations and Normative Implications

It seems that in the last few decades, rationalist models started giving way to emotive/intuitionist models of moral judgment. This came about due to the role of the empirical sciences in providing evidence for the role of emotions and intuitions in the making of moral judgments. However, even if the descriptive accounts favored the intuitionist's side, a normative account need not necessarily follow. Moral intuitions might be the main source of our moral judgments, but ought they be? Some argue that intuitions are a reliable source for moral judgments (e.g. Railton, 2014; 2017) while others argue they are not (e.g. Singer, 2005). Before we can answer this question, we need to identify what is meant by moral intuitions and what processes underlie them. In this paper I argue that our moral intuitions are not infallible, but not dumb brute gut-feelings either and that they can inform some normative moral theory. I back up my claim with findings from the literature.

# CONTENTS

ACKNOWLEDGEMENTS.....	v
ABSTRACT.....	vi
Chapter	
I. INTRODUCTION.....	1
II. MORAL INTUITIONS.....	6
A. Background.....	6
B. Haidt's Social Intuitionist Model (SIM).....	8
C. Railton on the Affective System.....	13
D. Reliability.....	18
E. Moral Foundations.....	22
III.OVERVIEW OF THE EVIDENCE.....	29
A. The Moral Brain.....	31
B. Findings from Experimental Philosophy and Psychology.....	42
C. Moral Reasoning.....	54
IV.LIMITATIONS AND MORAL EDUCATION.....	58
V. CONCLUSION.....	69
REFERENCES.....	70

# CHAPTER I

## INTRODUCTION

In an attempt to avoid falling for the naturalistic fallacy, many philosophers, for a long time, thought that the empirical sciences had little to offer moral philosophy. Simply any descriptive account that is sketched by developments in psychology and neuroscience regarding our moral behavior and beliefs should be irrelevant when discussing what we ought to do, for the normative does not follow from the descriptive. But even if that's the case, it's rather odd to claim that how we do act and how we are made to act by our psychology has no bearing on how we ought to act, for even if the latter does not follow directly from the former, as Sinnott-Armstrong (2006) puts it: "psychology can still affect moral philosophy in indirect ways". (p. 339). I don't necessarily claim by it that there's some sort of inferential relation between the psychological and the normative, but that understanding our psychology can help us advance a better empirically-informed moral philosophy. Before we delve into any normative implications, it's important, I believe, to see how our minds actually process moral situations and how we come about our moral judgments; specifically, what our minds can and cannot do when they entertain a moral dilemma. For it seems pointless to discuss what we ought to do if it turns out that we can't do it because we are limited by



our psychology<sup>1</sup>. In addition to that, it's important to make clear the foundations (if there are any) for our moral beliefs and behaviors and to see how they hold and whether they apply universally.

In recent years, philosophers started paying more attention to the great advances in psychology and neuroscience and the explanations they offered for why most people behave the way they do in certain moral situations. Intuitionist models were taking hold after rationalist models were dominating the moral philosophy debate. Under intuitionist models, our moral judgments are thought of as the outcomes of fast-acting intuitions rather than reasoned deliberations, as envisioned by rationalist models (see Haidt, 2001). Some attempted to anchor 'morality' in a set of foundations from which all of our moral judgments can be explained (see Haidt, 2012; Gray, Young & Waytz, 2012). Greene (2008) used evidence from neuroimaging to claim that: "[D]eontological judgments tend to be driven by emotional responses (...) rather than being grounded in moral reasoning" as opposed to Utilitarianism (i.e. consequentialism) which "involve[s] genuine moral reasoning" (p. 36). Greene's (2008) aim was to challenge the general belief that deontology is a rationalist model for moral judgments, and to shed light on the importance of taking the findings of psychology and neuroscience into account. This empirical claim, according to Greene (2008), "cast[s] doubt on deontology as a school of normative thought" (p. 36). Regardless whether his normative conclusion follows, the

---

<sup>1</sup> This point illustrates the "Ought Implies Can" (OIC) axiom, where it's generally agreed upon that a person can't be obliged to do something unless it is in her power to carry it out. If an action is humanly impossible, due to physical (or otherwise) constraints, then a person can't be held under a moral obligation to do it, nor be judged for not doing it (Mizrahi, 2015). This axiom is generally assumed and taken to be self-evidently true, but some argued that this might not necessarily be the case (e.g. see Kekes, 1984; Stocker, 1971). For the purposes of this paper, I will not delve much into this debate.

psychological basis of our moral theories, it is argued, does have an added value in our endeavor to find the correct, or most suitable normative moral theory<sup>2</sup>.

Before we can even begin to discuss the role of moral intuitions in philosophy, we need to give it a proper definition. There's little agreement on what exactly constitutes a moral intuition (or even just a general, non-moral intuition) in the literature, so much so that many appear to be arguing past each other. Some say intuitions are beliefs (e.g. Lewis, 1983)<sup>3</sup>, others claim that they are dispositions/inclinations to believe (e.g. Inwagen, 1997). One prevalent view takes an intuition to be "a *sui generis* occurrent propositional attitude, variously characterized as one in which a proposition occurrently *seems* true" (Pust, 2019)<sup>4</sup>. Those are some of the most prominent accounts regarding intuitions. Another point that should be made clear is whether moral intuitions significantly differ from non-moral intuitions, and if so, in what ways. If intuitions are beliefs, moral intuitions are moral beliefs (e.g. the belief that one ought to keep a promise they made) (e.g. Sinnott-Armstrong, Young & Cushman, 2010). On another account advocated for by McMahan (2013), "a moral intuition is a moral judgment (...)

---

<sup>2</sup> Even if Greene's (2008) normative conclusions (i.e. that deontology is not a rationalist theory, or an adequate normative theory) do not follow from the findings of his neuroimaging studies, it can still be argued that empirical findings are relevant for moral theory. It could be that Greene simply failed to articulate a valid argument or that some of the premises in his argument were not sound (which is what I will argue for later). We can still argue for the relevance of empirical findings irrespectively of Greene's success or failure in achieving his aim.

<sup>3</sup> Lewis (1983) says that "our "intuitions" are simply opinions".

<sup>4</sup> There are different, more discriminate variations of this that I will not discuss in this paper. It's enough to think of intuitions under this account as *sui generis* (unique) mental (conscious) states akin to perceptual/sensory experiences (Bedke, 2008, p. 253).

that is not the result of inferential reasoning” (pp. 104-105)<sup>5</sup>. There’s also a debate regarding the role of emotions in moral intuitions, where some argue that emotions are necessary, or both necessary and sufficient for moral judgments, while others reduce that role to one of motivation, where emotions follow from judgments (rather than influence them) and motivate action instead (Huebner, Dwyer & Hauser, 2008; 2009).

I view moral intuitions as system 1 generated moral judgments, where system 1 is a class of specific cognitive processes in the brain (that are unique to this system) and that issue in specific states (i.e. beliefs, judgments, desires, attitudes, responses, etc.) I rely on Railton’s (2014) account on the affective broad system – what he refers to System 1 by – and explain why this account best defends my claim that intuitions are more than mere brute gut-feelings. I aim in this paper to show that the features of the affective broad system, as explained by Railton (2014), are sufficient to provide support for viewing intuitions as ‘reliable’ sources for moral judgment<sup>6</sup>. I will go into detail explaining those features and show how they are aligned with the empirical evidence. I will also explain the dual-process system theory in psychology (which System 1 is part of) and introduce the two different accounts present in the literature, and show how they affect my view.

My goal in this paper is to stress the importance of intuitions in the making of moral judgments and the processes that underlie them. My stance is that intuitions are

---

<sup>5</sup> It’s immediate, and not inferred/does not follow from other prior beliefs. Some accounts equate moral judgments with moral beliefs, even though I believe it’s possible to have judgments that contradict one’s beliefs.

<sup>6</sup> I will flesh out what I take reliability to mean later on in this paper, but for now, when I say that intuitions are reliable, I do not mean to say that they are necessarily correct, or infallible.

judgments that, for the most part, are informed by reason, flexible and open to being attuned and educated, rather than being basic brute emotions or gut-feelings. Our moral theories (or any new ones we hope to sketch later) could make use of them or be informed by them. I also want to discuss, briefly, the importance of ‘Moral Learning’ and how it’s possible to attune our intuitions and how the affective system from which these intuitions emerge can allow for that.

## CHAPTER II

### MORAL INTUITIONS

#### A. Background

In philosophical tradition, if we take moral intuitions to have some normative authority, we mean that, for the most part, they “are true and [they] direct us to the discovery of foundational moral principles that are also true” (McMahan, 2013). Many philosophers adopt such a view, using our everyday intuitions as evidence (or data) from which they form a moral theory that best fits them, or attack theories that lead to judgments that contradict them. This is mostly evident in attacks against Utilitarianism, where our intuitions to some hypothetical thought experiments clash with major Utilitarian approaches to them. The most common thought experiments that are used to undermine the appeal to Utilitarianism are the footbridge thought experiment and the transplant case. In the transplant case, most people would immediately deem an act of killing one healthy individual to harvest her organs to save five an immoral one, an intuition that clashes with the philosophical theory of act Utilitarianism that emphasizes that an act is good if it optimizes well-being (Pust, 2019). This is also the case with the footbridge example, where most people would choose not to push a man off a bridge to stop a trolley, even if that meant saving five others (Singer, 2005, p. 340). These examples are ones “in which a philosophical theory is taken to be *prima facie* undermined by contradicting an intuition regarding a particular hypothetical case” (Pust, 2019). Intuitions such as those are taken at face value to be true and are used as evidence to either support certain moral theories or to put them down. Many philosophers don’t even realize that they are using them as evidence and completely

take them to be self-evidently true, in no need of justification themselves. This could be due to the appeal of some intuitions in which it does not cross someone's mind to doubt them. But should we not doubt them? Singer (2005) argues against the use of intuitions to undermine Utilitarian conclusions, saying that "we should be ready to challenge the intuitions that first come to mind when we are asked about a moral issue" (2005, p. 332). He argues against intuitionism, relying on evidence from neuroscience and evolutionary theory to drive his point that we ought to be more critical of intuitions and not take them to be true at face-value (Singer, 2005, p. 332).

Before we answer the challenges put forth by Singer and others, we need to address some of the misconceptions regarding the use of intuitions in moral philosophy that even many who subscribe to intuitionism seem to fall prey to. We need to address what moral intuitions are not. Even if we credit our intuitions with some normative authority (i.e. we deem them credible or trustworthy in leading us to the correct moral judgments), it's important to stray away from claiming that all of our moral intuitions are infallible or that we can never doubt them (McMahan, 2013). Moral intuitions are not a bullet-proof system that never fails, and claiming otherwise does a huge disservice to any normative account that wants to build on them. There are simply many instances that show our intuitions to be mistaken, or instances where our initial intuitions get rejected and change (McMahan, 2013). Add to that the fact that many of our intuitions were fostered in environments with cultural and moral norms that differ from those of

other cultures, thereby biasing our intuitions in favor of those norms<sup>7</sup>. Any normative account that involves moral intuitions should avoid falling for what some call ‘intuitive dogmatism’ for it will simply fail. Instead, we should identify how intuitions actually work and whether they are easily affected by morally irrelevant factors. We need to flesh out more the role of reason in moral intuitions and from there, we can argue as to whether moral intuitions can inform a moral theory or not.

### **B. Haidt’s Social Intuitionist Model (SIM)**

Haidt is arguably one of the most influential advocates for psychological intuitionism currently. Arguing against rationalist models that were dominating the discussion in moral psychology, he puts forth an alternative: “the social intuitionist model” (Haidt, 2001, p. 814). It’s a purely descriptive psychological model that is set out to explain how ordinary folk come up with moral judgments, and intuitionism is at its heart<sup>8</sup>. In this model, he asserts that people don’t come up with moral judgments through reasoning and reflection, but rather “that moral intuitions (including moral emotions) come first and directly cause moral judgments” (Haidt, 2001, p. 814). When a person is faced with a certain ‘eliciting situation’, a person’s already existing moral intuition is activated as a response to witnessing the situation, and a moral judgment follows as a

---

<sup>7</sup> Just consider the differences in intuitions of Liberals and Conservatives in the U.S. and how they were fostered by their surrounding environments (Haidt, 2012).

<sup>8</sup> By descriptive, Haidt wants to assert that his model is not meant to put forth any normative claims, departing from those who advocate for philosophical intuitionism. It’s simply about how people act and not how people ought to act (Haidt, 2001, p. 815). So his attack on rationalist models is of a limited sense, and it could be argued that his account could still be compatible with a normative rationalist model that would accept his descriptive intuitionist account while also arguing for a reason-based normative one.

direct cause from that intuition. For example, I come across a grown man striking a child in apparent anger. I immediately judge this action as wrong, influenced by a feeling of disapproval, anger, shock or whatever emotions such a situation might elicit in me. For Haidt (2001), a moral judgment<sup>9</sup> is an appraisal of an action or a person (as being good or bad, right or wrong, etc.) and it can sometimes be influenced (not inferred) by the culture and upbringing of the person making the judgment (p. 817). He (2001) also defines moral intuitions “as the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion” (p. 818). This account is similar to the one brought forth by McMahan (2013), where he equates moral intuitions with judgments and characterizes them as being non-inferential (they don’t follow from previous beliefs). Haidt (2001) also stresses the importance of affect in reaching such judgments, deeming our emotions vital to the whole process, and that’s due to him grounding our moral intuition in system one, or what he dubs the intuitive system, in the dual process theory of cognitive psychology. In sum, intuitions for Haidt are instantaneous “feeling[s] of approval or disapproval (...) with an affective valence” (Greene & Haidt, 2002).

But this account does not completely get rid of reasoning. It still plays a role in Haidt’s account, just not as much as it’s given credit for. Moral reasoning, for Haidt (2001), is a

---

<sup>9</sup> I want to make clear that throughout this paper I will be only concerned with judgments and not actions. When it’s said that intuitions directly cause moral judgments, I mean evaluations and not direct action. I do not mean to imply anywhere that an intuition directly causes someone to actually do something, for example approaching the man who struck the child in my previous example (though it can motivate someone to act).



conscious slow (i.e. consists of steps) cognitive process, it's "intentional, effortful, and controllable" and it "consists of transforming given information about people in order to reach a moral judgment" (p. 818). So if I want to reason my way towards a judgment regarding the man that struck the child, I should first take into account all the features of the incident and gather all relevant information before inferring a judgment. Since the process involves more steps, it's evidently slower and requires conscious processing. For Haidt (2001), this process is most prominently present after the judgment is made, not before, and he refers to the link going from judgment to reasoning as "the post hoc reasoning link" (p. 818). After the moral judgment is made, people sometimes feel inclined to justify their judgment so they use moral reasoning, not to reach their judgment, but to justify it (or rationalize it) after the fact. So if someone were to ask me why I judged the act of the man hitting the child as wrong, I could provide a list of reasons to justify my judgment, such as claiming that children are vulnerable, that adults ought to control their anger around them and so on. But these reasons did not necessarily cause my judgment, rather it was just an instant feeling of disapproval that did, and yet I claim that they were part of the judgment process. They amount to an attempt of rationalizing my judgment, nothing more.

However, Haidt (2001) allows for instances where moral reasoning can be used to reach moral judgments, either by using reason to over-step an existing intuition and reach a different judgment via the "reasoned judgment link", or by using reason to change the initial intuition directly, via the "private-reflection link" (p. 819). However, those links are rarely used and usually involve situations where we haven't already formed strong intuitions. We still get to the bulk of our judgments through the "intuitive judgment link" (Haidt, 2001, p. 819). In addition to that, through the social links in the model, a person's moral reasoning or judgment can be used to change or influence

another person's intuitions (Haidt, 2001, p. 819). For Haidt (2001), those links showcase the biggest role of moral reasoning, for since we are social animals, we need to form strong social circles, and so attempt to make allies and friends by getting them to think like us and therefore use reasoning to convince them of our perspective. But even the scope of reasoning here is limited<sup>10</sup>, and those most successful at inducing new intuitions are probably influential people (e.g. celebrities) who others look up to and want to imitate<sup>11</sup>.

Haidt's definitions of moral reasoning and moral intuition reflect the features of the "intuitive system" – system one – and the "reasoning system" – system two – in what has come to be known in cognitive psychology as the "dual-process system" (Haidt, 2001, p. 818)<sup>12</sup>. What Haidt's (2001) version of the dual-process system dictates is that reasoning and intuition belong to two distinct, independent cognitive "systems" (i.e. sets of processes) respectively, each with their unique set of characteristics. Intuitions are anchored in an evolutionarily ancient system that first evolved in our species to help us survive the dangers of our environment. This system is a system of automatic responses and reflexes, home to our impulses and emotions. It is holistic, fast, effortless, common to all mammals, and most importantly, the processes that underlie it

---

<sup>10</sup> Abortion and Capital punishment debates are good examples of how few actually change their minds when presented with moral arguments.

<sup>11</sup> For example, during the last primaries elections in the U.S., Taylor Swift, a popular singer, posted on social media urging young voters in her state to vote. The following day, a surge of young people got registered to vote (Wang, 2018).

<sup>12</sup> It's important to note that Haidt (2001) thinks that moral judgments, for the most part, are arrived at non-inferentially. Since he takes it that moral judgments are those generated by System 1, then it's implied that system 1 generated judgments are non-inferential whereas system 2 generated judgments are inferential. This distinction reflects his account and it's not a view that is necessarily shared by everybody (i.e. it's not a matter of definition that system 1 generated judgments are non-inferential, but rather a matter of Haidt's own view).

are non-accessible by us (Haidt, 2001, p. 818; Haidt, 2007). We can't identify the steps (if there are any) of how we reached a certain conclusion if the process took place in system One. No reasoning, no rational processing takes place in system one under this characterization. System two, on the other hand, takes care of all the rational conscious processes that we have. It's relatively new, in an evolutionary sense, and it is unique to us and possibly to some extremely intelligent primates. It's slow, conscious, involves clear multiple steps that are accessible to us, effortful and demands attentional resources (Haidt, 2001, p. 818; Haidt, 2007). Moral reasoning, unlike intuition, makes use of this system. For a long time, the dual-process system was understood in a manner similar to this, but recent research indicates that there might be more to the story, and that these two separate systems may actually depend more on each other than previously thought (Railton, 2014; Railton, 2017).

Haidt's social intuitionist model provides an interesting outlook on how people come up with moral judgments. The model itself fares well with the available empirical evidence and explains cases that Haidt dubs as "moral dumbfounding", where most people still insist open their initial intuitive moral judgments even if they fail to provide rational justification for them (Haidt, Bjorklund & Murphy, 2000). Where Haidt falls short, I argue, is in his definition of moral intuitions. He undermines the role of reasoning greatly and he bases his definition of moral intuition on an outdated dual-process system theory. He reduces intuitions to something akin to a 'gut-feeling'. Because of that, it would be hard to argue that intuitions can be of use in a normative theory of moral judgments as they are presented in Haidt's account. My view takes moral intuitions to be system one generated moral judgments, and if judgments are the product of this system as defined by Haidt, then our intuitions would appear in a very unfavorable light. In what follows, I will argue against Haidt by presenting another

account of system one put forth by Railton (2014) that challenges the view explained above and it's the one I advocate for in this paper<sup>13</sup>.

### **C. Railton on the Affective System**

Greene's (2008) findings were interpreted to mean that the moral theory of deontology is rooted in emotional responses. This descriptive psychological process the underlined deontology had certain normative implications. As Greene (2008) puts it: "our understanding of moral psychology (...) casts doubt on deontology as a school of normative moral thought" (p. 67). In other words, since intuitive emotional processes underlie deontological judgments, this undermines the normative appeal for deontological theories. But this last step does not necessarily follow from the descriptive work done before. For one, it's an assumption that intuitive and emotional processes are irrational and it's a claim that needs justifying. So many already assume that reason and intuition/emotion represent two opposing systems, but this might not be necessarily the case.

We tend to have a sort of pessimistic view regarding how our everyday intuitions come to be. We take them to be spontaneous, non-deliberative, not based on conscious reasoning, non-voluntary, unchanging and persistent even in face of reason, unsupported by satisfactory justification and yet, they have the power to move us to

---

<sup>13</sup> When I claim that intuitions are system-one generated judgments, I do not mean that all of the processes of system one are a matter of intuition. System one deals with automatic processes (such as the process of driving a car), acquired habits and with impulses and reflexes (such as recoiling from touching a heated object). System one deals with a wide range of automatic processes, only some of which are a matter of intuition.

make all kinds of moral decisions (Railton, 2014, p. 815). If painted this way, it seems like our intuitions can't be a reliable source for our moral judgments. We can't judge right from wrong based on a process that is not flexible and that rejects reasoning and involvement and change. However, Railton (2014) argues that the features of intuition highlighted above are nothing but "intuitions in the observational sense" (p. 815). In other words, it's how intuitions appear to us on the surface. He agrees that if this observational sense is all that there is about intuition, then they can't be counted as a reliable authority for making moral judgments (Railton, 2014, p. 815). But is this all there is to intuitions? The answer is no. According to recent studies in affective neuroscience, our affective system (i.e. system One), from which intuitions are realized, is much more flexible, and much more reliable, than first thought (Railton, 2014, p. 813). As Railton (2014) puts it: "intuitions [are] surface manifestations of an underlying structure of rules and information a tacit-competency-based model of intuition" (p. 817). In other words, intuitions seem to originate from a reliable and good source. And those intuitions that appear to be simple and spontaneous came to be by utilizing a complex multi-layered system that made use of both conscious and unconscious elements, skills and experiences, emotional intelligence, social influences and biases and so on (Railton, 2014, p. 822). Simply put, intuitions do not depend on a simple and 'stupid' non-cognitive, non-conscious underlying system, but rather on the workings of different systems and capacities that shape them into what they are now. These underlying complex processes do give a better understanding of how our intuitions come to be, and give them more credibility and 'intelligence' and reliability in many moral situations.

Railton (2014) then goes on to explain what is known in psychology as the dual-process model (p. 827). The dual process model speaks about the existence of two

separate, yet interacting cognitive systems, System One (or the broad affective system/the intuitive system in which processing is automatic, non-deliberative and not consciously accessible) and System Two (or the reasoning system, in which processing is effortful, intentional and more consciously accessible) (Haidt, 2001, p. 818; Railton, 2014, p. 827). Here, it's apparent that moral intuitions stem from system one, the affective system, whereas moral reasoning stems from system 2. Haidt (2007) views the affective system as "ancient, automatic, and very fast" and he contrasts it with "a phylogenetically newer, slower, and motivationally weaker cognitive system" (as cited in Railton, 2014, p. 828). But in this sense, the affective system is seen as one riddled with biases, dependent on blunt emotions, and one that reaches decisions through simplified mental short-cuts (Railton, 2014, p. 829). In other words, it does not appear as a reliable source from which important moral decisions should originate. But is the affective system really structured this way?

Railton (2014) gives a different perspective on the affective system, claiming that "it has come to be seen as a flexible, experience-based information-processing system quite capable of tracking statistical dependencies and of guiding behavioral selection via the balancing of costs, benefits, and risks" (p. 833). Even though its ancient, it is nevertheless evolving and is shaped to handle the adaptive needs demanded for the survival and flourishing of our species. This mechanism that evolved this way allows us to learn by utilizing our everyday life experiences. It's also receptive to outside information and capable of change, in addition to being able to contemplate risk-benefit scenarios that were originally thought to happen only on a conscious deliberative level (i.e. basic cost-benefit calculations could very well be automatic processes). These conclusions stem from findings in affective neuroscience that painted the affective system in a completely new light, by observing neural activity in the brain during the

process of decision making (Railton, 2014, pp. 834-835). These findings also move our understanding of intuitions closer to those classical notions of intuitions that were envisioned by philosophers such as Kant, who thought of intuitions as that which gives reasoning a steady foothold (Railton, 2014, pp. 825/834). As Railton (2014) puts it: “humans are able to act intuitively – “without stopping to think” (...) – in intelligent, organized, “plan like” ways that extend over time and are aptly responsive to a changing array of costs and benefits” (pp. 836-837). In addition to that, we have quite a lot of evidence to show that intuitions tend to change more frequently than we think they do. Just imagine how our attitudes regarding cases like abortion and gay marriage have shifted greatly in the last few decades. Those old attitudes have evolved and changed, mainly through social influence. We can refer to the links of social reasoning and persuasion in Haidt’s (2001) social intuitionist model that play an important role in shifting and changing the intuitions of others (p. 815). So intuitions are indeed more flexible and adaptive than we take them to be<sup>14</sup>.

This research that Railton (2014) presents tells us that the reason why we think of intuitions the way we do is because of a faulty understanding of the dual-process system theory. The old dual-process system theory, or what Railton (2017) refers to as the “first generation dual-process accounts of moral judgment” pushed the view that emotions/intuitions and reason belong to two distinct, independent systems. This created

---

<sup>14</sup> Railton (2014) gives a more descriptive definition of moral intuition, where he defines it based on its prominent features and origins in our cognitive system. He claims that our intuitions are anchored in what he dubs the “affective system”, which points out that emotions play an important role in intuitions. However, I don’t believe that this necessarily means that emotions are essential to every moral intuition. We can have abstract moral intuitions without an obvious affective element to them. It’s just that the same system that handles our emotional responses is also at the core of our moral intuitive ones.

a dichotomy that led many to believe that a process can't be both emotional/intuitive and rational, automatic and complex, fast and logical, highly motivational and flexible and so on. It's assumed that just because a process is emotional, or that it has a strong basis in emotions then that automatically disqualifies it from holding a rational status. Greene (2008) falls for this assumption when he argues that Deontological judgments can't be rational because they depend on emotions, as opposed to Utilitarian judgments that take into account cost-benefit considerations. But it's not very obvious whether this strict emotion/reason dichotomy is in fact true. At the very least, it's a claim that needs to be justified and argued for rather than just assumed true at face value. Maybe we ought to challenge the original dual-process system accounts that brought forth this strict dichotomy in the first place.

Recent research in cognitive and neuroscience paints a new picture of the dual-process system theory where this apparent reason/emotion dichotomy does not seem to be as pronounced as previously thought. Railton (2017) refers to these new accounts as the "second generation dual-process accounts of moral judgment". These new accounts are explained in terms of "the role of learning in moral psychology" (Railton, 2017). Basically they explain how we come to learn to judge and act in certain situations, where the equivalent of system one is responsible for acquiring and learning appropriate habitual responses and behaviors while the equivalent of system two is associated with goal-directed behavior and action (Railton, 2017). However, what's relevant here is that both systems are rational, flexible and allow for learning new responses and behaviors (Railton, 2017). Reason is not just a feature of system two, for system one makes use of it too. These systems, in other words, overlap, feed from each other and work together. In addition to that, their underlying neural networks and the "systems involved (...)" overlap extensively, and that each may play various roles in shaping how the brain uses



the other” (Railton, 2017). This is also the case when people make intuitive moral judgments as a response to moral dilemmas; as in intuition is not just the result of the sole work of an ancient, automatic and irrational affective system as previously thought. Railton (2017) uses this model to explain the apparent irrational discrepancies that we usually find in people’s intuitive replies to trolley problems<sup>15</sup>. He argues that their intuitions and answers are far from being irrational, but it’s that since we view our affective and reasoning systems from the lenses of the first-generation dual process accounts, we fail to make sense of the apparent asymmetry we find. This is a defense of intuitions, for if they truly make people consistent in their judgments, then they might be reliable. I’ll come back to Railton’s discussion on intuitions regarding trolley problems later on in this paper.

This newly adopted dual-process system theory, along with the features of moral intuitions that Railton describes give us a good reason to trust in our intuitions more. It’s those features of flexibility, openness to change and reliance on experience that are most essential to help us reach our reliability goal. It’s not enough to show that intuitions are more right than wrong, but it’s important to be able to pin-point where they can go wrong and whether there is hope to change them, or to educate them, so that they can become more trustworthy in the circumstances where they might err.

#### **D. Reliability**

---

<sup>15</sup> For the most part, people are likely to pull the lever in the trolley and loop cases but reluctant to push the man in the footbridge case.

The word reliable, when referring to intuitions, is a rather loaded word. I use it here in my paper because it's extensively used in the literature in describing intuitions that we can 'trust to be correct', though certain authors might mean different things by it (e.g. Demaree-Cotton, 2016; Railton, 2014, Alexander & Weinberg, 2014)<sup>16</sup>. For one, Alexander and Weinberg (2014) distinguish between two senses of reliability: the "*trustworthiness* sense of reliability" and the "*baseline accuracy* sense of reliability". To say that intuitions are reliable in the baseline accuracy sense, it simply means that intuitions are more likely to lead to correct judgments than not (Alexander & Weinberg, 2014). In other words, they are true more than half of the time. On the other hand, to say that intuitions are trustworthy is to say something over and above just that. It's not enough for this sense of reliability that intuitions are more right than not. What is needed is for us to be able to confidently know when intuitions fail specifically, not that they just do a certain percentage of the time (Alexander & Weinberg, 2014).

Alexander and Weinberg (2014) use an analogy with perception to make this distinction clearer, since many tend to think of intuitions as reliable in the same sense they think perceptions are (e.g. Audi, 2015). We generally trust that our senses are not deceiving us, however we are also aware that they err in certain unusual circumstances. For example, we can't completely trust our sight when looking at something very far away. But the fact that we can (for the most part) pinpoint the circumstances where our perceptions might fail allows us to know what should be done to overcome their shortcomings (e.g. we get closer to the object we want to look at). In this sense we can

---

<sup>16</sup> Railton specifically uses the word 'reliable' multiple times in his account on intuitions and the affective system and since I adopted his account, I felt inclined to mention that word as well.

keep thinking of our perceptual system as reliable (trustworthy) (Alexander & Weinberg, 2014). If it's the same for intuitions, we can say that they are trustworthy.

But it's this analogy with perception that is rather deceptive. There's so much talk about intuitions being reliable, yet reliable relative to what? And if we were to say that intuitions are biasing, biasing relative to what? Let's take visual perception: We deem sight reliable if the processes that underlie vision are themselves reliable. Those processes themselves are reliable relative to a generally-agreed upon objective criteria (taking into account the limitations of our scientific tools in discovering said criteria). A person has reliable sight, as in the object appears to her as it really is (assuming moral realism), if she is in the right conditions that allow her to see the object. So her visual cortex located in the occipital lobe is not damaged, her neural networks are not interrupted, her retina receives light as it's intended to, she's at an ideal distance from the object, there's sufficient light in the room and so on. Our knowledge of the workings of the visual system allow us to determine, to a certain degree of confidence, when our visual perceptions fail and when they don't. In that sense, we can speak of 'correct' visual perceptions. We can say things like person X is correct in saying that the chair is red when he looks at a red chair and that person Y is wrong in saying that this same chair is blue.

But this does not translate well with moral intuitions, not unless we properly determine the right conditions under which intuitions are judged to be correct. Some people have the intuition that same-sex relationships are wrong while others have the intuition that they are normal. Which one is correct, if we were to reject moral relativism, depends on the moral theory that is adopted by the speaker. So to speak of reliable intuitions in the sense of 'correct intuitions', we need to first introduce and defend a normative moral theory of judgment and to show why this specific theory is,

objectively, the right one. So for example, if we take deontology to be the correct moral theory, then ‘deontological’ judgments are the ones that are correct, and if we can determine that most intuitions lead to a deontological judgment, then we can claim that these intuitions are reliable. This would still be a hard claim to prove since we can’t really study all the intuitions that exist and map them out, keeping in mind that the methods being used in laboratories to study intuitions most probably do not reflect the moral situations that ordinary folks face all the time and their findings not necessarily generalizable to the real world (Alexander & Weinberg, 2014).

I am not advocating for a specific ethical theory in this paper nor am I sketching out a new one. Therefore, it would make little sense to speak of ‘correct intuitions’. My goal is to see if moral intuitions can inform<sup>17</sup> some normative moral theory and to do that, I try to prove that intuitions are, simply-put, not stupid. By relying on empirical evidence, I want to argue that intuitions can be rational (i.e. informed by reason), be the products of rational cognitive processes, consistent, flexible, capable of change and all the other features that Railton (2014) talks about. It is in this sense that I speak of reliable intuitions. In addition to that, when I talk about bias in intuitions, I won’t mean, in this limited sense, that which makes our intuitions err; rather, I will focus on the role of supposedly morally irrelevant factors (e.g. framing effects) and how they affect our intuitions. In other words, our intuitions are biasing if they are greatly influenced by morally irrelevant factors. I advocate for the position that our intuitions are not always

---

<sup>17</sup> Can add something of value, can contribute to or can point to certain normatively significant features in a moral theory. For example, Rawls’ theory of “Reflective Equilibrium” makes great use of moral intuitions.

helpless victims to such morally irrelevant factors as previously thought but they rather hold their ground pretty well (Demaree-Cotton, 2014).

### **E. Moral Foundations**

How do moral intuitions differ from regular non-moral intuitions? This is an important point that gets sometimes overlooked and its importance lies in the methodology that is used in studying said intuitions. For in a laboratory setting when we present our participants with a certain dilemma and inquire about their judgments, we need to make sure that we are actually studying what we are set out to study. If the questions we ask them do not actually reflect the way lay-persons perceive or think of morality, then we might not be actually studying their moral judgments but something else entirely. An example of this is the tendency for some researchers to conflate conventional with moral transgressions in studying moral judgments of their participants, unaware of the importance of this distinction in moral psychology (Blair, 1995). Intuitions about conventional transgressions may not necessarily be moral intuitions and so we need to ground moral intuitions in something and distinguish it from other intuitions<sup>18</sup>.

We have specific intuitions that are activated as a response to specific eliciting situations. So for example, a factual intuition is one that is elicited when we are asked to judge certain statements of fact as true or false (e.g. of a factual statement: H<sub>2</sub>O is Water). So a factual intuition to the statement “H<sub>2</sub>O is water” is the judgment that this

---

<sup>18</sup> I am claiming here that moral intuitions (i.e. system one generated moral judgments) are those that are activated as a response to specific moral situations, as in they are sensitive to specifically relevant moral features of the situation as opposed to other non-moral features.

statement is true. These intuitions tend to be affectively-neutral, as in they do not activate in us a strong emotional response. There are also kinds of intuitions that can be dubbed expert intuitions and these are intuitions that are acquired by a limited number of skilled individuals. So for example, a professional chess-player can have expert intuition when it comes to playing chess, where he can automatically perceive the most appropriate move to take down his opponent, or a doctor, who by just looking at certain symptoms exhibited by her patient, can immediately see something the rest of us can't (Pust, 2019). Similarly, moral intuitions are intuitions about moral situations, or what we perceive to be a moral situation. So for us to be able to accurately study moral intuitions, we need to figure out what counts as a moral situation to the lay-person. Several accounts were present in the literature whose aim were to determine the essence of morality in human nature. Two accounts that I find compelling and that are backed by a wealth of data are Haidt's (2012) Moral Foundation Theory and Gray et al.'s (2012) account on mind perception. It's important to keep in mind that both these accounts are only descriptive in nature and do not make any normative claims.

Haidt (2012) puts forth a model of morality that includes five different domains: harm, fairness, loyalty (or in-group), authority, and purity (p. 125). For him, these five represent the foundations of morality and his aim is to explain the discrepancies in people's moral judgments by appealing to these foundations, where some people tend to develop certain foundations more than others. So Conservatives in the United States, for example, are more likely than liberals to develop all five foundations, for they are more likely to view transgressions to purity, authority and loyalty foundations as wrong whereas liberals in general restrict such moral judgments to the foundations of harm and fairness (Haidt, 2012). So a violation of a loyalty foundation could be something like someone burning the flag of their country where it could be viewed as that person not

being loyal to their in-group. Someone who sufficiently developed the loyalty foundation is more likely to judge this action as wrong than someone who restricts her moral judgments to actions that cause harm, for instance. It's important to point out that all five foundations are present in everybody, as in these foundations are universal, and they are independent of each other. The difference is that some people develop certain foundations more than others, depending on their cultural and educational upbringing (Haidt, 2012). Now whether we ought to develop all five foundations or not is an open question that depends on more than just the present descriptive account. Haidt, however, seems to fall prey to the naturalistic fallacy by implying, through his theory, that it is best to reflect the entire intuitive system by properly developing all five domains and that no foundation takes superiority over another (Flanagan, 2014, p. 21)<sup>19</sup>. Going back to the descriptive claim, if Haidt's (2012) Moral Foundation Theory is correct, and that we truly view morality in terms of these distinct five foundations, then moral intuitions are intuitions specifically about harm, fairness, loyalty, authority and purity transgressions and those are the kinds of intuitions we should study in our experiments.

On the other hand, Gray et al. (2012) think of "mind perception" as the essence of morality. Mind perception entails perceiving or assigning minds (i.e. mental capacities of intention and/or experience) to other entities, being those other humans, supernatural beings, fetuses, dead people or your dog (Gray et al., 2012). Those minds might not exist; this does not matter as long as you perceive them to exist (Wegner & Gray, 2016).

---

<sup>19</sup> Haidt does not explicitly talk about what we ought to do, he even insists that his claim is purely descriptive (Haidt, 2012, p. 98). However, his normative account is implied in statements where he suggests that 'moral relativity/pluralism' is a good thing (Haidt, 2012, pp. 101-106)

Moral judgments are rooted in mind perception (Gray et al., 2012). In other words, whether you judge an action as right or wrong depends on your perceptions of the minds of the one doing the action (i.e. the doer) and the one receiving the consequences of the action (i.e. the feeler/experiencer). If the one doing the action is perceived to have the mental capacities of intention and thought then they are worthy of blame or praise, given that their actions affect some other mind perceived as capable of experience. All of morality<sup>20</sup>, including all of its different domains, simply fall into perceiving an interaction between two ‘kinds’ of minds: an agent – an intentional doer - and a patient – a feeler/experiencer (Gray et al., 2012). This forms sort of a cognitive template<sup>21</sup> – “the moral dyad” - from which all (or most) moral judgments can be explained (Gray et al., 2012). In addition to that, for an action to be deemed immoral, there should be perceived harm committed by a powerful agent on a vulnerable patient. Perceived harm<sup>22</sup>, therefore, is the criteria from which an action is judged to be morally wrong or right. And since morality is a combination of a perceived agent and a perceived patient, the more intentional and powerful an agent is perceived to be, and the more suffering a patient is perceived to endure, the more immoral an action becomes (Gray et al., 2012).

Unlike Haidt’s five foundations, Gray et al. (2012) claim that perceived violation of the harm domain is all that is needed to deem an action immoral. The moral

---

<sup>20</sup> Gray and others use Morality and Moral Judgments interchangeably. In this paper, I do not imply that they mean the same thing. I’m just explaining Moral Judgments from Gray and others’ perspective.

<sup>21</sup> It is hard to strictly define many concepts, so the human mind creates its own cognitive template, based on the characteristics of exemplars. Similar to how the word ‘dog’ brings to mind a four-legged creature with fur and a tail, even if not all dogs have fur or a tail. (Gray et al., 2012).

<sup>22</sup> What this indicates is that there is no need for the existence of actual objective harm. The mere perception of harm, even if there were none, is enough. In other words, without perception of harm, there won’t be a moral judgment.



dyad claims that all of Haidt's domains are mediated by perceived harm and fit into the dyad, even if they don't seem so at first blush<sup>23</sup>. So people see a situation as a moral one only if they perceive an agent harming a patient. It's important to point out that both Haidt and Gray and company's accounts stem from the belief that intuitions play the most crucial role in making moral judgments; it's just that Haidt (2012) claims that our intuitions expand over five different domains while Gray et al. (2012) take perceived harm as the sole foundation. So Moral intuitions under Gray et al. (2012) are concerned with perceptions of harm, but they also take into account perceptions of agency and experience in the individuals present in the moral situation. The moral actor sometimes gets ignored in moral psychology studies, where many focus on the morality of the action and deem the agent doing the action irrelevant to the participant's judgments, even though, under Gray's et al. (2012) account, the characteristics of the agent, their intentionality and their agency matter for the observer's judgment. Gray and Wegner (2011) claim that: "moral judgments are often more concerned with the people who complete acts rather than acts themselves". This is evident, for we judge a transgression committed by an adult more harshly than we do a child who carries out the same act, for example. Like Haidt's account, this one is also purely descriptive, and is not pushing the claim that we ought to take perceived harm as the sole criteria for moral judgments.

Both of these accounts try to anchor our moral intuitions in specific foundations and they present the empirical sciences with a starting point from which to study these intuitions. For the purposes of this paper, I will not argue in defense of one account over

---

<sup>23</sup> Violations of purity, authority and loyalty are seen as wrong, not because they are wrong in and of themselves, but because people view them as harmful (Gray et al., 2012).

the other but the review of the literature that I will be presenting below mostly fits better with the account of mind perception rather than with Haidt's (2012) moral foundation theory.

In addition to such accounts, the moral/conventional distinction should also be coined out, for participants' intuitions regarding a conventional transgression could differ significantly from their intuition regarding a moral one. A conventional transgression is an action or behavior that violates societal norms and rules and it's dependent on an authority deeming it impermissible. On the other hand, a moral transgression is concerned with violations to the welfare of others, is deemed more serious than a conventional transgression and is rule independent (Blair, 1995, pp. 5-6). For example, if the government decided one day to allow tax evasion, 'cheating' on one's taxes will no longer be seen as a transgression. However, moral transgressions like murder will not be okay if an authority figure permits them to take place, since they are wrong in and of themselves simply by being moral transgressions. Also, moral transgressions are generally seen as being more serious, and we advocate for harsher punishments to those that break them (Blair, 1995, p. 6). The reason for that is that moral transgressions are transgressions that cause harm to others, i.e. there's an obvious moral victim. The existence or perceived existence of a victim is enough to trigger in the bystander a sense of immense wrongdoing that can't be brushed away. This could happen by eliciting a strong emotional reaction to the violation – i.e. what Blair (1995) refers to as "moral emotions" (p. 4) – a reaction that can't be achieved equally by

violations of conventional norms. This distinction is an important one because it highlights the ways we judge transgressions and how not all transgressions are equally serious or equally worth our full time, energy and resources<sup>24</sup>. So our intuitions can react in vastly different ways depending on the kind of transgression we are presented with. It should be noted though that not everybody agrees with the above characterization and that it remains controversial whether such a distinction exists (Kelly, Stich, Haley, Eng, & Fessler, 2007).

In the next chapter I will present an overview of the research on moral intuitions, focusing on findings from neuroscience, behavioral science and experimental philosophy and psychology. The research focuses on the role of intuitions, emotions and reasoning in judgments, particularly utilitarian vs deontological judgments to trolley-like cases. In addition to that, evidence from neuroimaging highlights specific neural processes and regions of the brain that are associated with moral judgments, permitting us to see what kinds of processes underlie what kinds of judgments.

---

<sup>24</sup> It also might be the case that three of Haidt's moral foundations (i.e. loyalty, authority and purity) can be thought of as conventional transgressions rather than moral ones since they are mostly concerned with upholding certain norms and following rules.

## CHAPTER III

### OVERVIEW OF THE EVIDENCE

So far in the literature many experimenters were able to test folk intuitions regarding specific moral situations, trying to find a clear pattern of how people react to certain dilemmas compared to others. For example, Nichols and Knobe (2007) carried out philosophical experiments to test people's intuitions regarding moral responsibility and free-will, where their results indicate a tendency for people to be compatibilists when they are asked to judge a highly-affective concrete scenario (e.g. Is Bill morally responsible for raping and murdering a stranger in a deterministic universe?) and incompatibilists when asked to judge either an abstract condition (e.g. Ought we hold people morally responsible for their actions in a deterministic universe?) or low-affect concrete condition (e.g. Is Mark morally responsible for cheating on his taxes in a deterministic universe?). These results showed to them that our intuitions are inconsistent and that they are biased by the affective elements (which are supposedly morally irrelevant) present in some of the presentations of moral situations.

Others were interested in the debate between Deontology and Utilitarianism, arguing that although both were thought of as rationalist models, empirical findings were used to undermine the rational status of deontological judgments, claiming that they are anchored instead in emotional responses while utilitarian judgments are anchored in cognitive rational processes (Greene, 2008; Singer, 2005). Many also focused on a set of hypothetical dilemmas dubbed the "trolley problems", showing that people exhibit asymmetrical intuitions to different versions of these cases, even though, arguably the versions do not differ in any morally relevant features (Greene, 2008;

Singer, 2005; Nichols & Mallon, 2006). Some took this to show that intuitions simply can't be reliable and that we ought to "reconsider [their role] in normative ethics" (Singer, 2005) while others tried to reason away the apparent asymmetry, arguing that trolley-related intuitions are not as inconsistent as they appear at first blush (Railton, 2017; Everett, Pizarro & Crockett, 2016). In addition to that, many highlighted the features of certain moral dilemmas and the moral rules we subscribe to and the role they play in influencing our intuitions and reasoning (Greene, 2008; Nichols & Mallon, 2006; Tobia et al., 2013; Fernandez-Berrocal & Extremera, 2005; Valdesolo & DeSteno, 2006; Strohminger, Lewis & Meyer, 2010). Other researchers went back to moral reasoning and studied the effects of reasoning and reasoning manipulations on moral intuitions, indicating that we need not give up on the role of reasoning in moral judgments just yet (Sauer, 2012; Fine, 2006; Cushman, Young & Hauser, 2006; Paxton, Ungar & Greene, 2011; Paxton & Greene, 2010).

In addition to findings from psychology, neuroimaging studies helped map out the relevant neural basis for moral judgment processes in the brain. They identified the areas in the brain most active during the process of moral judgment while also differentiating the effects of emotion, intuition and reasoning in the brain (Greene et al., 2001; Kahane et al., 2012; Moll et al., 2002; Greene & Haidt, 2002; Cushman, Young & Greene, 2010; Moll, de Oliveira-Souza, Bramati & Grafman, 2002; Young & Dungan, 2012; Shenhav & Greene, 2014; Young & Koenigs, 2007; Pascual, Rodrigues & Gallardo-Pujol, 2013; Moll & de Oliveira-Souza, 2007; Funk & Gazzaniga, 2009; Cushman et al., 2012). In addition to that, they studied moral judgments in subjects suffering from damage in relevant brain areas to better map the role of these areas (Greene, 2007; Koenigs et al., 2007; Pujol et al., 2012). If we are interested in knowing the origins of intuitions in the brain, such findings prove fundamental to our purposes.

Greene (2008), for example, found that deontological judgments are carried out by emotional neural processes in the brain while utilitarian judgments are carried out by controlled cognitive processes. Kahane et al. (2012), on the other hand, asserts that Greene's findings don't reflect a deontological/utilitarian dichotomy but rather the brain regions he identified reflect the neural basis of intuitive vs counterintuitive moral judgments. The literature identified several features of moral judgments which I will talk about next.

### **A. The Moral Brain**

Are there specific regions in the brain that show increased activation when we process moral situations? Yes, studies that compared the brains of subjects using functional magnetic resonance imaging (fMRI) during the process of forming judgments in both moral and non-moral conditions highlighted increased activation in specific regions, such as the ventromedial prefrontal cortex (VMPC), among others (Young & Dungan, 2012). However, such findings don't necessarily indicate that this brain region is unique to moral cognition or that there is even a place just unique for it. For example, the VMPC is also implicated with emotional processing so it's not clear whether an increased activation in that area is responding to some affective element or to a specific moral component, assuming such a component exists (Young & Dungan, 2012). It appears that the process of moral judgment makes use of many other cognitive processes, some emotional, some rational, some both and so on. Add to that the role of

mental representations and theory of mind<sup>25</sup> in making moral judgments, indicated by the activation of the right temporoparietal junction (RTPJ), for we include perceptions of intention and agency of moral actors before we judge their actions (Young & Dungan, 2012; Gray et al., 2012). Then where would that leave us in our request to identify the moral brain – a unique brain region, or set of regions, or set of subsystems or neural processes, whatever it may be? Maybe the moral brain is everywhere – a part of the emotional brain and the rational brain and the social brain – or maybe it's nowhere for there might be no such thing as a unique moral element that is separate from all the other processes in the brain (Young & Dungan, 2012). To argue that intuitions are reliable, it might be helpful to figure out the areas of the brain that they originate from. Since intuitions are system one generated judgments, what are the brain areas that are involved in system one processes? And are some of these areas involved in rational processes, among others? If system one makes use of a limited section of the brain that is involved only with basic emotions and reflexes, like the amygdala for example, then it could be argued that our intuitions, that originate from such a system, are not reliable. Therefore, I think identifying the brain regions responsible for the workings of system one is helpful to our project.

Neuroscience is still ways off from answering the “where” question, however, that does not mean that the research that we have now does not have anything to add. So many studies tried to control for the influences of reason and emotion and other possible confounding variables in an attempt to try to understand how moral judgments get

---

<sup>25</sup> The Theory of Mind, briefly, is that which concerns itself with the minds and mental states of others and how we come to represent them and their beliefs and attitudes (Young & Dungan, 2012).

formed and their results highlight an important first step towards answering our question. In addition to that, even if they can't find a unique moral process from which our moral judgments depend on, by identifying which other processes are involved can also add to our understanding of how our moral judgments get formed, and whether these processes render them dependable. In the end our aim is to defend the claim that our moral intuitions and judgments are anchored in a reliable source which allows us to say that these intuitions can exert some normative authority in moral psychology.

A lot of the evidence on moral judgments from neuroscience and neuroimaging rest on a comparison between Utilitarian/Consequentialist and Deontological judgments regarding trolley-like problems (which are concerned with intended/foreseen physical harm)<sup>26</sup> (e.g. Greene, 2008). Greene (2008) was concerned with people's intuitions regarding trolley problems<sup>27</sup>. In the classical trolley dilemma, most people deem it morally permissible to pull a lever to direct the train to kill one person to save five. This judgment, Greene (2008) argues, is compatible with Consequentialism, which emphasizes that a good action is that which is concerned with maximizing well-being (i.e. the action with the best consequence). On the other hand, regarding the footbridge dilemma, people tend to have the opposite intuition, where it is deemed impermissible to push a man from a footbridge to stop a train heading to kill five people. This

---

<sup>26</sup> The research on moral judgments reflected in this paper greatly focuses on this specific moral transgression: explicit physical harm. The reason is because the literature available out there also predominantly focuses on explicit physical harm. This makes sense because there is a lack of consensus on what exactly constitutes a moral transgression, and since not many dispute that physical harm is one, using it would be a safe bet to ensure that one is actually studying moral judgments rather than conventional judgments. It would be interesting, nevertheless, for future research to explore more other kinds of moral transgressions.

<sup>27</sup> For detailed descriptions of the trolley problems, see Railton (2017)



judgment is dubbed a deontological judgment, according to Greene (2008), for it is one out of respect of the rule that forbids exerting harm on a person, regardless of the consequences of the action<sup>28</sup>. Greene (2008) hypothesizes that the distinction between the two cases is a distinction between personal and impersonal forms of harm. In the trolley dilemma, the harm is impersonal in the sense that all that is needed to save the five people on the tracks is to flip a switch, an action that holds little emotional weight. However, in the footbridge case, we need to physically push someone to their death, which is a very “up close and personal” thing to do and which carries a lot of emotional weight, mainly because we evolved aversive emotional reactions to these kinds of personal acts of violence (Greene, 2008). From there, he (2008) concludes that the essence of deontology must lie in our emotional responses, whereas consequentialism depends on the works of (non-emotional) cognitive controlled processes.

If that were the case, these findings should be reflected in neuroimaging results, where deontological judgments would show increased activation in brain regions associated with emotional processes while utilitarian judgments would show increased activation in regions associated with cost/benefit calculations (or more generally, regions associated with controlled processes) (Greene, 2008). In the case of deontological (i.e. personal) judgments, the brain regions that showed greater activation

---

<sup>28</sup> I find it unfortunate that Greene (2008) characterized these judgments in terms of consequentialism and deontology. For one, deontology concerns itself with more than just following basic rules, for it takes into account concepts of rights, autonomy and agency, rationality, duty and respect for persons, among others. Also, his definition of consequentialist judgments reflects, at best, basic tenets of act utilitarianism, which concerns itself only with immediate cost/benefit calculations. These are rather superficial definitions that do not do justice to the theories. Nevertheless, I will stick with these labels in my paper, mostly because the research I explain here makes use of them as well.

were: “the posterior cingulate cortex, the medial prefrontal cortex, and the amygdala” (Greene, 2008, p. 44). These areas are usually activated during processing of emotional stimuli, thereby fulfilling Greene’s (2008) prediction regarding the relation between deontology and emotions. Similarly, and as predicted, Consequentialist (i.e. impersonal) more judgments produced increased activation in “two classically “cognitive” brain areas, the dorsolateral prefrontal cortex and inferior parietal lobe” (Greene, 2008, p. 44). But what does that tell us about the origins of our moral judgments in the brain? Greene (2008) concludes from this that there are two distinct and independent psychological processes at work, where deontological judgments are rooted in emotional responses and utilitarian judgments are rooted in cognitive processes. So our moral intuitions that lead to such differing judgments make use of both systems, depending on the personal vs impersonal feature of the moral dilemma presented to them.

Why that matters, and why such findings lead Greene (2008) to conclude that consequentialism is a better normative moral theory than deontology, boil down to his personal/impersonal distinction. Whether an action involves physical contact or not is, arguably, morally irrelevant. So if our intuitions are very sensitive to the distinction as outlined by Greene (2008), then this challenges the claim that intuitions are reliable. If our intuitions are biased, and therefore inconsistent, by morally irrelevant features, then they are unlikely to be dependable. There are two ways to defend against this: We can either argue that the personal/impersonal distinction is a morally relevant one, or we can argue that it’s not this distinction that explains the asymmetry in our intuitions as a response to the trolley and footbridge examples. I will go back to this later on in my paper when I discuss the findings of experimental philosophy and psychology.

Going back to Greene’s (2008) fMRI findings on the making of intuitive judgments, it’s important to unpack the significance of the activated regions in the

brain. First, the role of the amygdala was not as prevalent as previously thought in affectively-salient intuitive judgments, but rather a heightened activation “was observed in the visual and left premotor cortex” (Kahane et al., 2012). This might be due to the fact that those areas are concerned with empathetic activation, an emotion associated with role-taking, whereas the amygdala is more concerned with basic emotions such as disgust (Kahane et al., 2012)<sup>29</sup>. This might reflect a distinction between basic emotions and unique moral emotions as well. On the other hand, counterintuitive judgments reflected a significant activation mostly in the rostral anterior cingulate cortex (rACC), indicating that counterintuitive judgments reflect only a specific kind of controlled cognitive processes since it generates increased activation in a much more limited area than what was previously thought (Kahane et al., 2012)<sup>30</sup>. The rACC is mostly concerned with cost/benefit calculations and emotional conflict resolutions, but it is also activated during feelings of guilt and when we consider what other people think of us (Kahane et al., 2012). This activation makes sense since those who make counterintuitive judgments are usually breaking certain moral rules and are probably aware of that, hence the feeling of guilt and hence why processes concerned with conflict resolutions get activated. But the increased activation of the rACC is not enough to show that counterintuitive judgments actually do rely on moral reasoning, at least not fully, for the areas that are mostly concerned with problem solving did not

---

<sup>29</sup> This gives credence to Gray’s et al. (2012) mind perception theory where it is argued that morality depends on assigning minds and mental states to other individuals.

<sup>30</sup> Controlled cognitive processes associated with activities such as complex problem solving are usually present more dorsally and there was no significant great activation in that area during the processing of counterintuitive judgments (Kahane et al., 2012)

show significant increase in activation (Kahane et al., 2012). These findings do not support the conclusion that overcoming one's intuition is the more rational thing to do. In other words, counterintuitive judgments (e.g. pushing the man off the footbridge) are not necessarily more rational than intuitive judgments, since they don't show significant activation in brain areas related to controlled processes. And intuitive judgments are not necessarily basic, or 'stupid', since they also lead to activation in areas that deal with more than just basic emotions<sup>31</sup>.

Many intuitive accounts of moral judgment associated moral emotions with basic emotions that can be activated in non-moral affective situations. If we consider moral intuitions as nothing more than gut-feelings, an instant activation of a basic emotion upon encountering a moral situation, then we would conclude that moral emotions – emotions activated as a response to specifically moral situations – do not differ from basic emotions – emotions that are activated as a response to everyday non-moral affective situations, such as disgust upon seeing a very dirty toilet. Haidt (2001) claims that moral intuitions are nothing over-and-above gut-feelings, which is a problem if we want to argue in favor of a moral theory that makes use of intuitions, since it's not very convincing to argue that we ought to trust in our instant basic gut-feelings. However, evidence from neuroimaging points out to differences in the processes that underlie moral emotions relative to basic, non-moral emotions (Moll et al., 2002).

---

<sup>31</sup> This does not necessarily mean that intuitive judgments that show limited activation are irrational. What I wanted to show was that system one depends on a much wider range of processes that in turn activate a significant number of brain areas relative to these processes. This supports Railton's (2014;2017) account on the affective system and how it works alongside the processes of system two as well.

Results indicate that while moral and non-moral unpleasant emotions share some neural substrates (i.e. “the amygdala, upper midbrain, right thalamus, superior colliculus, extrastraite visual cortex, temporal and coronal slices”), specific moral stimuli invoked further “increased activation of the right medial OFC and the medial frontal gyrus (MedFG) and the cortex surrounding the right posterior temporal sulcus (STS)” (Moll et al., 2002)<sup>32</sup>. As Moll et al. (2002) put it, these results indicate that: “the effects related to moral stimuli cannot be explained on the basis of the levels of emotional valence or visual arousal alone” (p. 2733)<sup>33</sup>. So there’s evidence that shows that moral emotions, and thereby moral intuitions in emotionally-charged moral situations, go beyond simple gut feelings.

The areas implicated in moral emotions are concerned with interpersonal relations, social schemas and representations of the mental states of others (Moll et al., 2002). Specifically, increased activation in the OFC reflect processes involved in “decision making related to reward and punishment” along with empathic processes (Moll et al., 2002). The STS, on the other hand, is associated with “more complex social cognition mechanisms” that go beyond basic emotional responses. In general, these areas were linked “to the development and maturation of moral, social and emotional behavior” and go beyond just reacting to present physical attribute of the stimuli (Moll

---

<sup>32</sup> It should be noted that the experiment deployed by Moll et al. (2002), from which these findings were extracted, focused on visual stimuli, hence explaining the activation in regions associated with processing visual stimuli. Other studies that focused on moral vs non-moral statements showed activations in regions associated with processing propositional statements (Young & Dungan, 2012).

<sup>33</sup> An example of a stimuli that is meant to induce negative moral emotion is an image portraying physical assault whereas an example of stimuli that is meant to induce negative non-moral emotion is a picture of body lesions or dangerous animals (Moll et al., 2002).

et al., 2002). These findings can offer support to the new dual-process system theory that claims that moral intuitions make use of both the emotional system one and the cognitive “rational” system two, since these two systems do work together and overlap (Railton, 2017). Based on such findings, I think it’s safe to move on from equating moral emotions to blunt gut-feelings and to admit that they undergo much more complex and cognitive processes<sup>34</sup>.

Another interesting branch of research in neuroscience focuses on individuals who suffer damages to certain brain areas. Prominent examples of such individuals are psychopaths, but there were also studies carried out on patients suffering from damage to the prefrontal cortex (e.g. Koenigs et al., 2007). Psychopaths were always of interest to behavioral researchers since they don’t necessarily do poorly in moral reasoning evaluations but yet are notoriously unmoved by the horrible (or anti-social) judgments that they make. Blair (1995) attributes that to their inability to distinguish between conventional and moral transgressions due to their lack of ability to feel empathy or to represent the mental and emotional states of other people. Another study tried to see how psychopaths fared compared to controls on judgments that reflect Haidt’s (2012) five moral foundations, and found that significant differences were only found with respect to the harm and fairness foundations, but not the others (Aharoni, Antonenko & Kiehl, 2011). These findings might also reflect the conventional/moral distinction, where the harm and fairness foundations correspond to the moral whereas the others (respect for authority, in-group loyalty and purity) correspond to the conventional. Pujol

---

<sup>34</sup> For a comprehensive review on the findings of neuroimaging studies, see Greene and Haidt (2002).

et al. (2012) studied the brains of psychopaths during the process of making moral judgments and realized that they “were more likely to endorse harmful action” and that they “showed significantly reduced activation in the medial frontal cortex and the posterior cingulate cortex and additionally in both hippocampus and posterior-inferior midbrain” (p. 919). These areas are mostly associated with emotional processing (specifically processes related to fear from punishment) and goal-directed reasoning, in addition to the lack of activation in dorsolateral prefrontal cortex (areas usually associated with problem-solving) which indicates that psychopaths don’t reach judgments by invoking cognitive rational processes either (Pujol et al., 2012). Psychopaths were always used to endorse the claim that moral intuitions and emotional processing are essential to moral judgments and that without them, reasoning enough wouldn’t be compelling to get individuals to reach the correct moral judgments. The finding that psychopaths also showed a lack of activation in areas related to controlled processes might indicate some sort of correlation between damage to emotional processes and more cognitive controlled ones. Maybe we need to readdress the claim that psychopaths don’t struggle with reasoning, at least when it comes to moral situations.

Regarding studies that targeted individuals with damage to the VMPC (an area involved in the generation of social emotions), researchers found that VMPC patients are mostly utilitarian and that damage to the VMPC allows them to consider judgments as a simple cost/benefit equation (Koenigs et al., 2007). Such patients don’t perceive a conflict between emotional moral rules that are strictly against moral violations like harming others and the more utilitarian judgment of increasing welfare, hence why they tend to take the utilitarian route in moral dilemmas such as judging the act of pushing a man from the footbridge more permissible than healthy patients (Greene, 2007). This

finding led Greene (2007) to claim that “patients with emotional deficits may, in some contexts, be the more pro-social of all”, since they mostly care about increased welfare. I find this association between utilitarian judgments and more rational cognitive processes that Greene (2008) advocates for rather problematic. Simple cost/benefit calculations are not necessarily rational, or rather the product of system 2 processes, for these too can be automatic<sup>35</sup>. So just because VMPC patients make judgments solely based on cost/benefit calculations does not in any way imply that they rely on the activation of controlled processes, or that they are free from the influences of morally irrelevant factors that affect the intuitions of people with an undamaged VMPC.

It’s important to remember that so much goes into making a moral judgment and that the consequences of an action alone are not all that matters, for we are also concerned with the intentions of the agents. Agency and intention matter for we are able to distinguish between intentional actions and accidental actions, giving more moral weight to the former. For people who are unable to represent intention or agency in others, they might end up judging accidental harm as more morally wrong than acts of unsuccessful harm, simply because they judge them exclusively on their consequences. This was observed in individuals with low activation in the right temporoparietal junction (RTPJ), an area associated with mental representations of other people, where these individuals gave a lesser role to intention in their judgments (Young & Dungan, 2012). Relying on just cost/benefit comparisons could lead us to judge children, for example, more harshly for moral transgressions than we ought to do, and make us

---

<sup>35</sup> For instance, Railton (2014) argues that system one of the dual-process system theory does take into account cost/benefit analysis.



ignore other morally-relevant features of moral judgments such as agency and intention<sup>36</sup>.

## **B. Findings from Experimental Philosophy and Psychology**

Many experiments were carried out to identify the intuitions of lay-persons as a response to certain moral dilemmas. Behavioral results reflected a tendency for intuitions to change depending on different factors such as a distinction between personal and impersonal harm (e.g. Greene, 2008), concrete vs abstract framing (e.g. Nichols & Knobe, 2007), manipulations of emotional context (e.g. Valdesolo & DeSteno, 2006), order framing effects (e.g. Demaree-Cotton, 2016), first-person vs third-person framing effects (e.g. Tobia et al., 2013), and intentional vs foreseen harm (e.g. Cushman et al., 2006). The problem with having our intuitions changing as a response to such factors is that some of them are morally irrelevant and to have them exert such influence on intuitions could make us doubt the reliability of intuitions in making moral judgments. In other words, if our intuitions are heavily sensitive to morally irrelevant features, then it could endanger the view that they are reliable. In this section I will try to argue that this is not the case and that our intuitions can guard well against the influence of morally irrelevant elements of moral situations.

Framing effects – the way a situation is framed where certain morally irrelevant features are manipulated by experimenters to study their influence on responses – have

---

<sup>36</sup> This is not to claim that Utilitarian theories do not take into account agency and intention, for what matters are the intended consequences and not the accidental ones.

been shown to, in certain cases, significantly alter people's intuitions (i.e. lead to different moral judgments depending on how the situation is framed or presented). This, again, is a threat to the view that intuitions are reliable. Tobia et al. (2013) compared the intuitions of philosophers to those of lay-persons and found that both groups fell victim to the "actor-observer" framing effect<sup>37</sup>, even though the results were in different directions. They (2013) use this finding to suggest that expert intuitions might be different from folk intuitions, but not necessarily better since they also fall victim to framing effects. What matters here is that the "actor-observer" framing effect<sup>38</sup> is a morally irrelevant feature, since it would seem unlikely for things to be permissible to only me but not to others (i.e. if an action is deemed impermissible/wrong for some stranger X, it should also be the case that this same action be impermissible/wrong for me) (Tobia et al., 2013). I argue, however, that this framing effect is not a morally irrelevant feature of moral judgment. As argued above, moral judgments make great use of representing and assigning mental states to others. Grey et al.'s (2012) mind perception theory highlights the importance of the role of the agent, their capacities and motivations, in the making of moral judgments. In other words, the actor (i.e. the agent) matters for our moral judgment. Features relating to agency, intention, capacities such as strength and influence and power the agent has, motivations and so on are all taken into account when making judgments, and hence replacing the agent in a moral situation

---

<sup>37</sup> In the actor condition, participants are given a vignette in which they are the person carrying out the action and are asked what *they* should do (e.g. *You* come across an injured person, is it morally permissible for you to not tend to their injuries?), while in the observer condition, participants are given a vignette about some other person X and are asked what *person X* should do (e.g. *Person X* comes across an injured person, is it morally permissible for them to not tend to their injuries?).

<sup>38</sup> Also most commonly known as the first-person/third-person framing effect.

is a morally relevant change. So in regards to the participants in the above study, having the unknown actor be replaced with the participant allows the participant access to a host of information, for she then takes into account information regarding her capacities, her moral theories and attitudes, her emotional states and so on, features and information she could not infer from the unknown actor in the “observer” condition (Demaree-Cotton, 2016). So the differences between how philosophers and lay-persons reach moral judgments might be relevant beyond the framing condition, since they differed.

In addition to that, Demaree-Cotton (2016) argued that the influence of framing effects (i.e. morally irrelevant factors) is exaggerated for the findings don’t actually reflect such a significant impact of framing effects on intuitions. She (2016) analyzed a great deal of framing effects studies and found that: “the probability that participants would *not* have expressed a different belief had the frame been different is relatively high, clustering predominantly in the eighties and seventies to give an average of 80%” (p. 9). In addition to that, the studies that showed the highest framing effects were regarding the “actor-observer” framing effect, and we have reasons to doubt that this framing is morally irrelevant, as argued above (Demaree-Cotton, 2016). In other words, our intuitive judgments are well guarded against the effects of framing effects that can be thought of as morally irrelevant. Our moral judgments hold their ground and are not easily changed or biased by a manipulation of a framing effect.

But it could be argued that the 80% figure is not enough and that the fact that people are affected by framing effects 20% of the time still poses a threat to the view that intuitions are reliable. However, there’s a difference between experimental moral scenarios and everyday moral dilemmas, where there is evidence that the influence of framing effects diminishes greatly in the latter, beyond the 20% average. Take for

example moral situations of which we have strong opinions about, such as abortion or capital punishment. It seems very unlikely that a framing effect will influence our judgments to such issues. Everyday moral scenarios that we face are not as ambiguous or as unusual as the ones studied in labs, and hence we might already have strong opinions about them and that would shelter us more from framing effects (Demaree-Cotton, 2016). In addition to that, individual differences in characteristics can mitigate the influence of some framing-effects, making some more or less susceptible to them (Demaree-Cotton, 2016). Fernandez-Berrocal and Extremera (2005) found that participants who had higher scores on emotional intelligence (EI)<sup>39</sup> were less likely to fall for framing effects. In addition to that, age was also found to have a significant effect, where university students fared better than high school students on the presented tasks, indicating that maturity (i.e. accumulated life experience/ fully-formed cognitive brain regions) can also shield us from the biasing influence of framing effects (Fernandez-Berrocal & Extremera, 2005). This makes sense since we take intuitions to be highly informed by experience. Future empirical research needs to tackle those differences and account for them in future studies, for it would do good to be able to identify them and be aware of them as we make moral judgments<sup>40</sup>.

---

<sup>39</sup> The EI scale used in this study addressed (1) Attention (i.e. the tendency to think of one's emotions and moods), (2) Clarity (i.e. the ability to distinguish between feelings and moods) and (3) Repair (i.e. the tendency to regulate one's own feelings) (Fernandez-Berrocal & Extremera, 2005). The reason why EI was deemed important in moral decision making is because "EI involves a striking balance between emotion and reason in which neither is completely in control" (Fernandez-Berrocal & Extremera, 2005).

<sup>40</sup> For example, people with higher verbal and mathematical SAT scores were less likely to fall for framing effects (as cited in Fernandez-Berrocal & Extremera, 2005).

Another arguably morally-irrelevant factor that might bias our intuitions and that might explain the asymmetry in our intuitions regarding the trolley and footbridge dilemmas is the personal/impersonal distinction (Greene, 2008). So Greene (2008) proposes that we are reluctant to push the man from the footbridge even if that meant saving five people because this action is a personal and violent one and much more emotionally salient than the action of flipping a switch, which is characterized as being impersonal. This principle of “contact-harm”, where harm achieved through physical contact with the victim is judged as worse than harm achieved through no physical contact, was also cited as justification by some participants for their moral judgments (Cushman et al., 2006)<sup>41</sup>.

However, others have argued against the personal/impersonal distinction, claiming that it’s “overly crude and unable to explain the variability found in responses to the trolley problems” (McGuire, Langdon, Coltheart & Mackenzie, 2009). As in, it doesn’t matter whether the personal/impersonal distinction is a morally relevant one or not, for it’s not this distinction that participants’ intuitions to trolley-problems are influenced by. Other versions of the trolley-problem underscore the personal/impersonal distinction even further. Everett et al. (2016) introduced a dilemma similar to that of the footbridge one, but instead of using physical force to push the man off the bridge, participants would flip a switch that would activate a trapdoor that would drop the man from the bridge and onto the incoming train. The trapdoor case does not fit Greene’s

---

<sup>41</sup> In this case, even though participants admitted that their moral judgments were influenced by the contact principle, they were unwilling to endorse it since they thought it wasn’t morally relevant (Cushman et al., 2006).

(2008) characterization of personal dilemmas, and yet, most people in the trapdoor case endorsed the deontological judgment (i.e. claimed it impermissible to activate the trapdoor) (Everett et al., 2016). Many other counterexamples also threaten Greene's Personal Hypothesis, like specific cases involving acts of self-defense and punishment that can be very personal and emotional and yet judged permissible nevertheless (Nichols & Mallon, 2006). To sum up, Greene (2008) is mistaken in claiming that the personal/impersonal distinction is what explains our differing intuitions. So how do we integrate all of these differing intuitions to the trolley problems?

Some argued that it's the Doctrine of Double Effect (DDE) that is responsible for the difference in intuitions between the footbridge case and the trolley case (Greene, 2008). Briefly, the DDE concerns itself with a distinction between intended harm and foreseen harm/harm as a side-effect, where the former is seen as morally impermissible while the latter as morally permissible. Since we use the fat man on the bridge as a means to save the other five, the harm is intentional and therefore judged to be morally impermissible. On the other hand, in the trolley problem, we don't intend to use the man on the side-track as a means to save the five, thereby rendering his death as a foreseen side-effect only which is judged by the DDE as morally permissible (Cushman et al., 2010). Assuming the DDE is a valid moral principle, could our intuitions be truly responding to it? Cushman et al. (2006) found evidence that people do employ the DDE in their judgments, however they do so unconsciously for they are not able to cite it as a justification. Irrespectively of that, there is some evidence that points to our intuitions being sensitive to distinctions between intended and foreseen harm. This distinction is a morally relevant one and so it won't endanger the view of our intuitions being reliable if they were to be sensitive to such a distinction.

A certain variation to the original trolley problem might put this conclusion at risk though, and that's the loop case. In the loop case, the person is used as a means to stop the trolley from killing the five individuals on the main track, and yet most people think it morally permissible to pull the switch (Greene, 2008). But certain different versions of the loop case, controlling for the intentional/foreseen harm distinction, shows a small, but significant effect in support of the DDE. Cushman et al. (2010) cite a study that compared responses to two different versions of the loop case: the first is the original loop case where the man on the side-track is used as a means to stop the train while the next is a case where a wall on the side-track is what stops the train but there's a man standing in front of the wall who will die as a foreseen side-effect to pulling the switch. The results showed that "subjects reliably judged the looped means case to be morally worse than the looped side effect case" (Cushman et al., 2010). The DDE highlights an important morally-relevant Kantian feature, which is respect for rational persons. According to Kantian theories, we ought not to use rational persons as means to an end. The DDE might not account for all of the differences in our intuitions in the trolley dilemmas, but it's a better candidate than the personal/impersonal distinction. Again, since the DDE reflects a morally-relevant feature, our intuitions being sensitive to it does not deem them unreliable. Even further, it could be argued that our intuitions are rational in being able to respond and change according to such a feature.

Another important factor that might play a role in explaining the apparent asymmetry in our intuitions to footbridge and trolley-like cases is trustworthiness (Railton, 2017; Everett et al., 2016). Railton (2017) conducted an experiment with the students of his undergraduate class and noticed that the asymmetry of responses in footbridge and trolley (or switch) cases were reflected in two other variations of the trolley-problem, mainly the cases of Wave and Beckon. Briefly, in the wave case,

participants were asked whether one should wave for the five workers on the tracks to move to the right to avoid the incoming trolley even if that meant another person on the left of the track will also see the wave and interpret it as being beckoned to step on the tracks, resulting in him dying. Contrast that with the beckon case, where participants were asked whether one should beckon a large man to step on the tracks so that his body can stop a trolley heading towards five workers (Railton, 2017)<sup>42</sup>. The responses of the students mirrored those of the trolley/footbridge responses, where the majority said that one should wave in the wave case but one should not wave in the beckon case (Railton, 2017). It's important to note that unlike the footbridge/trolley case, the wave/beckon case did not involve a distinction between personal and impersonal harm since the action was done at a distance, and yet, the asymmetry still holds.

Railton (2017) then asked his students to rate the people on trust depending on the answer they gave to the proposed dilemmas (e.g. if you knew that your roommate voted to push in the footbridge case, would you rate them as more or less trustworthy?). In the wave and trolley cases, students reported no change in levels of trust, however, in the beckon and footbridge cases, students reported a significant reduction in trust in people who chose to push and beckon in these cases (Railton, 2017). This result even held when accounting for the student's own choices, as in students did not just simply trust those that agreed with them (Railton, 2017). Citing evidence from the literature for certain characteristic traits of people who are disposed to answer push to the footbridge case revealed a higher score on psychopathy scale, along with "decreased levels of

---

<sup>42</sup> For more detailed descriptions of the wave/beckon cases, see Railton (2017).



empathy, harm-aversion, and perspective taking” (Railton, 2017). Those are rather undesirable characteristic traits and it seems that participants in such studies are aware of this. Even when presented with a purely hypothetical situation and even when participants are not urged to put themselves in place of the agent, many of them spontaneously do that, thereby simulating the situation in their minds and taking into account consequences along with representations of their own states of minds and those of others. In addition to that, participants are also concerned with how the act will appear to others and how they would feel doing such an act or how they would be made to justify it if asked to afterwards (Railton, 2017). All of these processes happen spontaneously at the exact moment the decision is being made.

By going through this process, Railton (2017) says that participants find that “simulating pushing the man “feels” aversive, “looks” surprising and callous, “seems” as if it would be hard to defend” (p. 14). This is not the case for the trolley problem and many participants find that people who do not react in this aversive way to pushing a man from the bridge exhibit an “underlying affect and motivation system [that] is displaced from the normal range”, something akin to moral indifference rather than moral altruism which in turn invokes feelings of distrust towards such people (Railton, 2017).

We could grant that the results reflect a tendency to distrust people who choose push or beckon but then what normative relevance does that have to morality? Should the good or right thing to do depend on such characteristic traits or on whether a person is trusted as a member of a social group? Maybe the right thing to do ought to be done even if it results in general distrust by the public. However, there’s evidence that characteristic traits and levels of trust play an important normative role in moral theory. For one, the tradition of moral philosophy includes theories that stressed out the

importance of virtuous characters and not just doing virtuous acts (e.g. Virtue ethics) (Railton, 2017). This means that agents matter, what they think matter and what motivates them to act matter. It's a mistake to dissociate actions from the agents carrying them out and to claim that "optimal individual acts" should take precedence over "good social relations" which would factor more into the common good (Railton, 2017). As Railton (2017) puts it: "Most human goods, and especially such life-sustaining goods as friendship, family, social solidarity, mutual respect, and humane caring, depend not only upon the acts performed but upon the attitudes, affect, and will of those involved" (p. 15). We depend on our social circles for survival and we are therefore sensitive to their emotional and motivational states and to their good and ill-will towards us and other people, and people who are willing to push others or beckon others to their death reflect a certain kind of ill-will and urges us to reject the idea of a society composed mostly of people who think like that (Railton, 2017, p. 15). So there's a good argument out there for the moral relevance of trustworthiness. Therefore, our intuitions to the series of different trolley-like problems are not random and inconsistent, and don't necessarily depend on morally-irrelevant factors. It's the opposite, they appear rational.

Railton's (2017) study did not conform to the usual standards in psychological experiments and did not control for other independent factors and this might deem his results invalid. However, Everett et al. (2016) were able to replicate most of the findings in Railton's philosophical experiment by conducting five thorough and controlled psychological experiments. Everett et al. (2016) were concerned with perceptions of trust towards participants who made utilitarian (push in the footbridge case) vs deontological judgments (not to push in the footbridge case) and found that most people trusted those who made utilitarian judgments less and were less likely to take them for

social partners or to trust them in economic games. These findings held even when researchers accounted for the participants' own judgments, so that the results were not explained by participants just trusting those they agreed with (Everett et al., 2016). The results of this study were attributed to people generally trusting those who express more characteristically deontological traits, like "respect for persons and commitment to social cooperation" (Everett et al., 2016).

Most studies around moral intuitions focused on judgments of permissibility to trolley-like cases. Straying away from this tradition, a philosophical experiment carried out by Nichols and Knobe (2007) focused on intuitions about moral responsibility in hypothetical deterministic universes in an attempt to figure out whether these intuitions were compatibilist in nature or not. Their findings revealed that people's intuitions regarding moral responsibility were mediated by several confounding variables, such as affect as well as the abstract or concrete framing of the presented moral questions (Nichols & Knobe, 2007)<sup>43</sup>. Agents in highly-affective concrete cases were deemed morally responsible for their actions even in a deterministic universe, whereas agents in low-affective concrete cases and in abstract cases were not held morally responsible for their actions in a deterministic universe only, but they were held responsible in an indeterministic universe (Nichols & Knobe, 2007)<sup>44</sup>. These findings show that people's

---

<sup>43</sup> An abstract framing of a moral question could be something like this: Should we hold individuals morally responsible for their actions? Whereas an example of a concrete framing would be something like this: Should Bill be held responsible for murdering his wife so that he can continue his affair with his secretary?

<sup>44</sup> The high affect condition in their study was about a man named Bill who stalks and rapes strangers, whereas the low affect condition was about a man named Mark who cheats on his taxes (Nichols & Knobe, 2007).

compatibilist intuitions depend on the affective valence of the scenario which pushed Nichols and Knobe (2007) to conclude that affect biases intuitions and renders them unreliable. These results seem to conflict with my view, indicating that affect biases our intuitions. However, I think that Nichols and Knobe (2007) missed an important confounding variable that might have affected their findings, and I will explain that below.

I disagree with their conclusion because I believe they ignored the influence of other confounding variables in their study. For one, they did not account for the conventional/moral distinction when it came to the transgressions that were outlined in their examples. Their high affect condition (i.e. Bill stalks and rapes strangers) is an obvious moral transgression, for there's a clear identifiable victim, a clear identifiable aggressor/agent and clear harm being carried out by the agent onto the victim. So participants will be quick to perceive this scenario as a moral transgression. However, the low affect condition (i.e. Mark cheats on his taxes) is not that clear. For one, there's no obvious identifiable victim or obvious harm being carried out against anyone. This case may as well be also a moral transgression, but there's no way to make sure whether the participants of Nichols and Knobe's (2007) study actually perceive it that way. If they perceive Mark's case as a simple conventional transgression, then they will treat it as such, thereby judging it less harshly and having no strong moral intuitions towards it. A conventional transgression is dependent on breaking a certain rule that forbids it, whereas a moral transgression is one that is wrong independent of any rules, so it's no surprise that people are reluctant to give it a pass, even in a hypothetical deterministic universe. Whether that's the case or not is an empirical question so future studies need to account for the conventional/moral distinction.

In addition to that, there's evidence that people have compatibilist intuitions regarding moral transgressions that lack a strong affective component, such as cases of petty theft. A similar example was illustrated in Nichols and Knobe's (2007) own review of the empirical literature on this matter, where they discussed the case of a man named Jeremy Hall who robs a bank in a hypothetical deterministic universe, and found that most people (83%) held Jeremy fully morally responsible for the theft (p. 667). Theft is another clear case of a moral transgression because it involves an agent (the thief), a victim (the person being robbed) and harm (the act of robbery), yet many cases of theft do not necessarily evoke in us strong affective responses. In conclusion, it's rash to judge the role of emotions as one that biases our intuitions and lead to us to make inconsistent judgments based on the findings of this study.

### **C. Moral Reasoning**

Haidt (2001) argues in his "social intuitionist model" that the role of reasoning in the making of moral judgments is very limited. Although I agree with his general premise that moral intuitions are the ones that mostly drive moral judgments, I disagree with his view on the influence of moral reasoning on judgment. Haidt (2001) mostly speaks of moral reasoning in terms of post hoc justifications (i.e. after the fact) and concludes that the reasoning people supply for their judgments play no causal role whatsoever in the moral judgment process. In other words, under this account, moral reasoning is nothing but a matter of confabulation (i.e. making up or fabricating reasons that played no actual effective role in reaching the judgment and claiming that they did) (Sauer, 2012).

However, as Sauer (2012) points out, just because moral reasoning is post-hoc, that does not necessarily mean the reasons supplied did not actually play a role in the making of moral judgments. Also, just because a procedure is automatic does not mean it's not

rational, and just because we don't have conscious access to the reasons that underlie our intuitions the moment we make judgments does not mean that these reasons did not play an effective causal role in the process (Sauer, 2012). So we can grant that most reason is post hoc without going away with reasoning's causal role in reaching moral judgment.

To illustrate this point further, Sauer (2012) discusses a common (non-intuitive) automatic action that many of us do daily: riding a bike. Riding a bike from work to home is mostly an automatic habitual process that does not rely much on attentional resources. The whole process is automatic, from unlocking the bike to the paddling motion to taking the road towards home to parking the bike. Yet it would be odd to claim that these automatic processes are nothing but a pointless series of actions without any reason or purpose. There is a purpose, mainly getting home, and if the rider is not always consciously aware of the reasons why they are riding their bike does not mean the action is not rational and that the reasons do not causally influence the action. An action being automatic does not render it irrational or not based on reasons (Sauer, 2012). Similarly, our intuitive process being automatic does not immediately render it devoid of reason<sup>45</sup>. Of course, not all post hoc reasoning is correct, some are genuine confabulations, especially in cases where people refuse to change an irrational conviction even when they can't find proper justification for it anymore (Sauer, 2012).

---

<sup>45</sup> Going back to Haidt's (2001) example regarding incest, it would appear hasty to claim that the justifications participants gave as to why they judge incest to be wrong a matter of mere confabulation. Reasons such as the risk of pregnancy are rational ones that might as well have played a role in why this particular action is judged wrong. The reasons need not necessarily be consciously present during the process of judgment, but they could have played a causal role in reaching said judgment nevertheless.

What we are arguing against here is equating all post hoc reasoning with confabulation (Sauer, 2012).

The other way some studies attempted to figure out the causal role of reasoning is by manipulating subjects to reflect and reason more before making their moral judgments (e.g. Paxton et al., 2012). Paxton et al. (2012) found that by inducing subjects to be more reflective (through letting them complete a Cognitive Reflective test before introducing the moral dilemma), they were able to increase the frequency of Utilitarian responses. In another study, they recorded the influence of argument strength and deliberation time on moral judgments surrounding incest and found that strong arguments proved more persuasive than weak arguments, but only when participants had more time to deliberate and reflect (Paxton et al., 2012). This finding illustrates that good arguments can be persuasive enough to push people to voice approval for emotionally-charged taboo cases such as incest. It also proves the importance of giving people the much needed attentional resources that reasoning requires to take place (in this experiment, that's sufficient time). This also challenges Haidt's (2001) moral dumbfounding claim that people are very unlikely to change their minds to cases like incest even if they are presented with good arguments. Even in some of Haidt's own work on moral dumbfounding, we can find the influence of certain individual characteristics on moral judgment. For example, higher socio-economic status (SES) participants tend to judge harmless taboo actions (e.g. eating one's dead dog) as less morally wrong than lower SES participants, suggesting that levels of education (higher SES participants tend to be more educated) do affect our moral intuitions (Sauer, 2012).

The role of moral reasoning remains uncertain and more studies need to be carried out to identify its significance in moral judgment (Paxton & Greene, 2010). However, my aim in this chapter was to represent a rather comprehensive review of the

literature and to try to identify the relevant factors that shape and influence moral intuitions. I focused on the role of morally relevant and irrelevant features in shaping our intuitions, arguing that our intuitions are not as influenced by morally irrelevant features as previously thought. This strengthens my view on intuitions that deems them reliable since they are not easily affected by morally irrelevant factors. I also tried to illustrate how intuitions are more consistent and rational than they first appear to be, focusing on discussions surrounding folk intuitions to trolley problems. Much of the evidence remains inconclusive, yet leaves space for a rather positive outlook on the role of intuitions in moral judgment. The processes that underlie our intuitions and the behavioral responses that result from them support our aim in showing that intuitions are indeed informed by reason and experience and indeed indispensable in any future normative moral theory that we might conjure. However, there are certain limitations that should be addressed and ways to answer to those limitations, which I will briefly mention in the final part of this paper.



## CHAPTER IV

### LIMITATIONS AND MORAL EDUCATION

There are situations that elicit incredibly high affective reactions in participants and they sometimes lead us astray. For example, the immense anger and fear that many have felt after the 9/11 attacks lead to an increase in anti-immigration sentiments and was used to justify increased and more restrictive security measures<sup>46</sup>. It appears that affect can bias our intuitions in highly affective situations and influence our moral judgments. Specifically, if the person themselves are the victim of a moral transgression, they are more likely to be influenced by their emotions into not making the best judgment. This is precisely why we argue that independent objective judges tackle cases of harm for we can't trust to make the right choices if we are the person being aggrieved in such a situation.

The most extreme cases of abhorrent moral judgments are reflected in studies in the psychology of atrocity (Doris & Murphy, 2007). Doris and Murphy (2007) argue that conditions of war can be defined as “cognitively degrading circumstances” that chop away at people’s cognitive capacities. People subject to such extreme conditions lose their ability to reason and reflect properly, to interpret situations properly and to

---

<sup>46</sup> Schüller (2016) found that 9/11 attacks lead to an increase in negative attitudes towards foreigners and immigration among the German public (as in the attacks had effects that went beyond US borders), though interestingly, she found that levels of education mediated the effects of the 9/11 attacks on such attitudes. This gives credence to the claim that our intuitions can be ‘protected’ in a way from the effects of certain unusual or highly-affective situations.

assign mental representations to others properly (Doris & Murphy, 2007). This is evident in how some American soldiers who participated in atrocities in the Vietnam and Iraq wars did not see the Vietnamese and Iraqis as people anymore, or in Gray et al.'s (2012) terminology, couldn't assign minds and mental states to them anymore and this might have justified in their minds the atrocities they carried out on them<sup>47</sup>. Doris and Murphy (2007) make a strong claim based on this, saying that "perpetrators of atrocity typically occupy excusing conditions and are therefore not morally responsible for their conduct" (p. 26). Whether they are right in saying that such people ought not to be held morally responsible is up for debate, but what's clear is that certain conditions, especially highly-affective conditions that elicit feelings of extreme fear and a context of violence and focus on following orders and group loyalty influence our intuitions to make horrible moral judgments<sup>48</sup>. Even if our intuitions are reliable in everyday contexts, should we be content with them even if they fail in such unusual circumstances?

---

<sup>47</sup> For example, Steven Dale Green, an American war veteran who was sentenced for raping a 14-year-old Iraqi girl and killing her along with her parents and sister said that "he didn't think of Iraqi civilians as humans after being exposed to extreme war zone violence" (Barrouquere, 2010).

<sup>48</sup> Doris and Murphy (2007) focus in their analysis on war on moral behavior (or action) rather than judgments and intuitions. So it could be said that behavior is compromised due to the 'damage' to the cognitive and affective systems of soldiers. It's also not clear that the behavior is caused by a certain intuition or judgment. However, the conditions that these soldiers behave in affect system one extensively, and since system one is at the core of intuitions, it could be that they end up being compromised as well. It's clear in the literature that some soldiers form certain attitudes about the people they carry out atrocities on during wars (e.g. how they think of the 'other' as not human anymore) and these attitudes can stem from the violence and the other conditions that they are exposed to during war (e.g. seeing their comrades getting killed can trigger an immediate negative attitude towards the group they are fighting against). I argue that such attitudes (they might not necessarily cause the morally reprehensible behavior but could very well motivate it) arise automatically, as in soldiers react to the death of their comrades, to the loud noises of war, to the putrid odor of rotting flesh and so on rather than them reasoning their way to such conclusions. These attitudes lead to judgments such as 'it's not wrong to torture prisoners of war' or 'enemy soldiers and civilians are the same'. Such judgments can motivate the behavior that these soldiers end up carrying out.

I find the study of moral behavior in war interesting because it highlights how supposedly ‘normal’ people – i.e. neurotypical people who do not suffer from severe mental illness or brain damage – can carry out abhorrent acts in highly-affective conditions such as war even though they would never act in such a way back in their homes and neighborhoods. If we take their behavior to be the product of intuitions and if our intuitions can be compromised to such a degree under specific situations, then it can be argued that they are not reliable. Or rather since such conditions compromise our affective system (i.e. system one) in general and since this system is at the core of our intuitions, then by extension our intuitions can also be compromised. So if immense anger or immense fear can bias our intuitions to such an extent, what stops anybody from falling off the rail when subject to such highly-emotional conditions?

There’s not much we can do in such highly-charged environments. For one, such conditions also impair reasoning and other controlled cognitive processes, so a rationalist approach won’t fix the problem. The problem lies with the cognitively degrading conditions themselves and it’s a wake-up call to try to improve such conditions and to change some aspects of military culture. We can also talk about ‘strengthening’ people’s intuitions so that they can be more prepared to deal with such situations. For intuitions are flexible and upon to learning and change and there might be certain tactics in place that can train people’s intuitions to be more resilient to the conditions people are forced into, but there’s not enough insight right now on how this can be done.

Rationalist approaches in war-like situations might not be very viable. Aside from the claim that reasoning is also negatively impacted in highly-charged and violent situations, reasoning also requires attentional resources. Our attentional resources (i.e. amount of attention, focus, time, control, effort etc.) are limited and can be directed

efficiently towards only a few tasks at once. A soldier that is worried about an ambush, trying to concentrate amidst loud war noises, concerned with following orders or carrying out a task by a commander on time cannot be expected to employ any more attentional resources in a situation that requires fast responses. When soldiers have seconds to decide what the best course of action is to avoid dying, they won't be able to reason their way to the best outcome; they would have to act immediately, thereby relying mostly on the automatic processes of system one. Therefore, if we want to reduce the acts of atrocities carried out by soldiers, the best course of action is to 'train' or 'educate' their intuitive system, thereby strengthening it against the influence of war-like conditions. More research needs to be done on what kind of 'education' should be carried out but it's a step to shift our attention to the workings of system one again, because for the most part, it's this system that guides our judgments, and probably motivates our behavior.

A good analogy, I think, regarding the workings of system one in highly-affective or stressful situations, can be made with the case of skilled emergency room (ER) doctors or trauma surgeons during a great influx of patients. Assume a big accident takes place, for example a train accident, that leads to a great number of injured individuals in need of immediate medical care. The few doctors on call need to be able to tend to many individuals at the same time, and many might be bleeding and in need of quick interventions. In other words, the ER doctor can't afford to ponder over every patient and try to reason her way to the best plan going forward. She needs to immediately make a decision, to prioritize the patients that need her care the most, delegate her work properly and so on. This situation is a stressful one and her decisions hold a lot of weight to them since a wrong decision can mean that a patient might die. But the immense cognitive load brought about from such a situation depletes her

attentional resources and therefore she has to rely on her intuitive and automatic system to make decisions. If such stressful situations truly impair system one greatly, then relying on it would be catastrophic. And yet, good ER doctors, using their expert intuitions that are informed by years of experience, manage to do their job well, even under unusual and extreme conditions. Their training allows their intuitive system to function reliably in such situations. Similarly, if we can find a way to properly train or educate our affective system to respond to highly-affective moral situations, then we might be able to shield our intuitions from their adverse effects.

Another limitation that was pointed at in the literature can be traced back to one specific emotion: Empathy (Bloom, 2017). Moral intuitions make great use of empathy, as is evident above from the findings of neuroimaging studies and from the general claim that people deploy empathy in role-taking and representing the mental states of others. In addition to that, people generally associate empathy with good effects, like pushing us to help others and better understand their plight. However, it's just this property of empathy that might be its downfall when it comes to our intuitions. Bloom (2017) defines empathy as "putting yourself in other people's shoes, feeling what you think they are feeling" (p. 24). And even though "empathy can motivate good actions", it "has a dark side" (Bloom, 2017). According to Bloom (2017), empathy does not work on large scales for we can't be expected to put ourselves in the shoes of millions of suffering people around the world. This will lead to cognitive overload, cognitive burnout and fatigue. Empathy then can lead us to ignore atrocities with high statistical causalities and push us to focus more on individual cases that involve one or few identifiable victims that we can relate to (Bloom, 2017). Also, empathy can clash with fairness (e.g. we give way too many resources to help one identifiable victim even if there are many others who are in more pressing need for them), be used to advocate for

animosity towards other groups and be susceptible to biases such as prejudice towards race and physical attraction (Bloom, 2017). Empathy also can't accommodate cases with no clear identifiable victims so it can't be used to influence people to act against certain harmful governmental policies (Bloom, 2017). Bloom (2017) then argues that this renders empathy a "poor guide to moral decision-making" (p. 27). Slovic and Västfjäll (2010) found similar results when they studied how people reacted to great losses of life, arguing that our intuitions in general are insensitive to them and "mislead us in the face of natural disasters or human disasters associated with poverty, disease, and violence". Even though they grant that our intuitions are in general reliable, it's just that they face shortcomings in this specific area, and they discuss the failure of a particular emotion, mainly compassion (Slovic & Västfjäll, 2010)<sup>49</sup>. In other words, empathy is an emotion that biases our intuitions, as in it renders them unreliable by making them insensitive to arguably morally relevant features (such as great loss of life).

But it's not clear that empathy is the culprit here. To say that empathy biases our intuitions in cases involving great losses of life is to say that we experience empathy in such cases and that this specific emotion pushes us to be insensitive to such moral situations. Since empathy is an emotion that involves role-taking and that is best equipped when dealing with a single (or few) identifiable victim(s), it could be argued that when faced with situations that do not involve such victim(s) nor is it possible to

---

<sup>49</sup> Though it might be the case that Slovic and Västfjäll (2010) failed to differentiate between empathy and compassion in their study since Bloom (2017) argues that compassion accounts for much of the shortcomings of empathy in cases that involve a great loss of life.

properly take the perspective of others, empathy is not even present. It's a specific emotional state that is brought up as a response to specific situations. So what Bloom and others take to be a failure of empathy is just situations that don't even evoke this emotion in the first place. It's not that empathy biases our intuitive responses to such cases, it's that empathy is not even present, and if it's not present, it can't bias anything<sup>50</sup>. Cases that deal with great loss of life usually evoke emotions like sympathy that don't involve putting ourselves in the shoes of others. Research can tackle this question better by making studies that can identify the exact emotions participants experience when presented with such cases. We can even make use of the strengths of an emotion like empathy and frame our moral dilemma in a way that appeals more to it. For example, Slovic and Västfjäll (2010) suggest we use images of victims that people can actually identify with rather than just use statistics, thereby drawing people's attention to the cause we are advocating for. Empathy can push us to do so much good, we just need to know when it's ideal to depend on it and how to frame situations to best monopolize that good.

Moreover, our intuitions may indeed fail, and that's okay. I never attempted to show that they are infallible, just that they are reliable, and they are indeed reliable in everyday cases under usual circumstances. However, I also stress the importance of knowing *when* intuitions fail so that we can address those failures and propose solutions for them. Studies as the ones above highlight some of the circumstances in which this happens and being aware of those circumstances allows us to combat them better. I

---

<sup>50</sup> This point regarding empathy was pointed out to me in a discussion with professor Quinn Gibson.

argued that Railton's (2014) account of the affective system grants our everyday intuitions flexibility. In other words, they can change and they can take into account new information and apply them accordingly. Intuitions can also be 'educated' so we can for example employ certain training to strengthen our intuitive responses to cases where they usually fail. Intuitive responses also make use of reason, so we can employ controlled cognitive processes to reduce the influence of biases on our intuitive judgments. The claim that sometimes certain morally irrelevant features can influence or bias out intuitions does not doom our account. This account advocated for here rests on the idea that our affective system, from which moral intuitions are generated, can guard well, to a certain degree, against such influences and is flexible enough to allow for improvements and learning so that we can counter their effects better when they are presented again in the future.

There's also evidence that people are aware of when their intuitions fail in certain contexts. Sauer (2012) talks about a process called "situation selection", in which people either remove themselves from a situation they know would bias their judgments, or expose themselves to situations that would address and reduce their biases (p. 267). Let's say I'm on a diet and I know that I have a weakness to pastries and that there's a pastry shop on a street that I usually frequent. Knowing that I'm weak-willed to such a temptation, I decide to avoid the street all-together, thereby removing myself from a situation that might lead me to cheat on my diet. Similarly, people can be aware of their prejudices and weaknesses in a moral context and take steps to either avoid situations that might elicit such prejudices or even take steps to fix them by putting themselves in situations that will directly treat those prejudices. For example, people who are aware of their implicit racist attitudes can take steps to engage people of different races or take classes on diversity so as to get rid of or reduce the influences of



such attitudes (Sauer, 2012). Again, it's the flexible nature of our intuitions that allows them to change and address the biases that might plague them.

Another limitation is one that addresses the methodology of the studies on moral intuitions. For example, Kahane and Shackel (2010) argue that studies in the neuroscience of moral judgments, specifically the neuroimaging studies carried by Greene, are insensitive to many confounding factors, mainly those relating to ascribing states to participants. In addition to that, the worry stems from the examples researchers use to study moral intuitions. Maybe reliance on trolley dilemmas does not reflect the intuitive processes of everyday moral judgments and maybe the items used in research do not study what they are supposed to study<sup>51</sup>. For example, the conventional/moral distinction is not very clear in the literature and many researchers are insensitive to it (e.g. Nichols & Knobe, 2007). Even the conventional/moral distinction as defined by Blair (1995) is not accepted by everybody (e.g. Kelly, Stich, Haley, Eng, & Fessler, 2007). Maybe the problem is a conceptual one. Throughout this paper I discussed a comprehensive overview of the literature and throughout, the concepts of emotions, intuitions and judgments were used interchangeable to sometimes refer to the same

---

<sup>51</sup> In a laboratory setting, participants are usually presented with a detailed moral scenario, where all the information of the case are explicitly explained. But in real-life scenarios, we can't immediately tell all the features of a moral situation. We might need to deduce certain features by observing the agents and actions taking place. For example, the intentions of an agent might not be explicitly announced so we might have to deduce those; the experience of the victim, how they are feeling, the harm they are being subjected to, etc. might also be something that's not very clear and depend on the observer. It might not even be clear what the cost/benefit ratio of a certain outcome is. So in the trolley dilemma, the cost/benefit calculation is clear: You pull the switch, one person dies and five survive. But certain real-life situations, the observer might not have all the information needed to make an accurate calculation. In other words, the way our intuitions function in real-life cases don't necessarily reflect the way they function in controlled environments such as a laboratory. We perhaps need to stray away from trolley-like cases and try to come up with situations that mimic our everyday life.

thing. This is not because I think that moral emotions are equivocal to moral intuitions or judgments or that I think emotion mediates all our intuition, it's more because the research itself uses these concepts that way. Different researchers define intuitions differently and many can't seem to agree on one comprehensive definition, thereby running the risk that experiments are not even studying the same thing. Future studies need to be aware of these problems in methodology and account for them.

There's hope that those limitations can be addressed and that we can attune our intuitions to respond better in contexts that might lead to bias. The processes that underlie our intuitions are rational, flexible and open to learning, despite their automatic nature. Due to this, we can speak of something called 'Moral Education'. Sauer (2012) and Railton (2017) argue that we can educate moral intuitions. Railton (2017) points out "that much of our learning takes place by observing others rather than through direct external reward or punishment". Our intuitions make use of experiences and they can learn through them. In addition to that, Sauer (2012) argues that our moral intuitions can be educated through direct explicit teaching as well and that its effects can be found in even small children. He (2012) claims that children can be made to understand the conventional/moral distinction depending on the reasons parents offer as to why each transgression is wrong.

Sauer (2012) links the process of moral education to habituation, where repeated usage of cognitive controlled processes become automated in time, so that they can be integrated into our automatic intuitive system to be used later in an effortless-perception like manner. Fine (2006) brings up a similar case regarding the act of stereotyping. So people who try to avoid applying their embedded stereotypes of others in certain situations use controlled processes to stem the effects of their prejudiced or stereotypical attitudes that get activated upon exposure to such situations. So the

stereotype that ‘black men are aggressive’, for example, might get activated when I come across a black man or read about a case that involves a black man. Since I’m aware of the stereotype and I want to put an effort to counter it, I try to avoid applying it since I feel it’s unacceptable to do so. If I keep using a conscious effort to stop myself from applying the activated stereotype, at some point, through the process of habituation, this conscious process gets automated. As in, I no longer need to consciously avoid applying the stereotype, I just automatically do (Fine, 2006). This process, Fine (2006) argues, also leads to the person no longer even activating the stereotype, not just not applying it. Since I repeatedly used conscious effort to challenge my activated stereotype as to not apply it, through this repetition, the whole process becomes automatic and the stereotype might no longer be activated upon encountering situations that used to elicit it before.

The field of moral education is a wide and relatively new area of research that shows great promise. Whether there’s some sort of criteria for ‘ethical education’ that is similar to other fields of education is not clear yet and goes beyond the limited scope of this paper. Philosophers and moral psychologists should for now move on from the rational-intuitive debate and put it to rest, focusing rather on whether the idea of moral learning is feasible, and if so, how do they go about with it.

## CHAPTER V

### CONCLUSION

Railton (2014) best sums up the affective system as “a system designed to inform thought and action in flexible, experience-based, statistically sophisticated, and representationally complex ways – grounding us in, and attuning us to, reality” (p. 846). The intuitions that have their source in such a system are therefore credible and a reflection of a wide and complex system that works most of the time. Such intuitions can be relied on to inform a moral theory in philosophy, despite some limitations.

## REFERENCES

- Aharoni, E., Antonenko, O., & Kiehl, K. A. (2011). Disparities in the moral intuitions of criminal offenders: The role of psychopathy. *Journal of Research in Personality*, 45(3), 322-327. doi:10.1016/j.jrp.2011.02.005
- Alexander, J., & Weinberg, J. M. (2014). The ‘unreliability’ of epistemic intuitions. *Current controversies in experimental philosophy*, 128-145.
- Audi, R. (2015). Intuition and its place in ethics. *Journal of the American Philosophical Association*, 1(1), 57. doi:10.1017/apa.2014.29
- Barrouquere, B. (2010, December 19). Ex-soldier talks about slaying of Iraqi family. *NBC News*, Retrieved from [http://www.nbcnews.com/id/40739938/ns/us\\_news-crime\\_and\\_courts/t/ex-soldier-talks-about-slaying-iraqi-family/#.XRrwAegzbIU](http://www.nbcnews.com/id/40739938/ns/us_news-crime_and_courts/t/ex-soldier-talks-about-slaying-iraqi-family/#.XRrwAegzbIU)
- Bedke, M. S. (2008). Ethical intuitions: What they are, what they are not, and how they justify. *American Philosophical Quarterly*, 45(3), 253-269.
- Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, 57(1), 1-29. doi:10.1016/0010-0277(95)00676-P
- Bloom, P. (2017). Empathy and its discontents. *Trends in Cognitive Sciences* 21 (1): 24-31.
- Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2012). Judgment before principle: Engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience*, 7(8), 888-895. doi:10.1093/scan/nsr072

- Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford handbook of moral psychology*, 47-71.
- Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12), 1082-1089. doi:10.1111/j.1467-9280.2006.01834.x
- Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology*, 29(1), 1-22. doi:10.1080/09515089.2014.989967
- Doris, J. M. & Murphy, D. (2007). From my Lai to abu ghraib: The moral psychology of atrocity. *Midwest Studies in Philosophy* 31 (1):25-55.
- Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology. General*, 145(6), 772-787. doi:10.1037/xge0000165
- Fernandez-Berrocal, P., & Extremera, N. (2005). about emotional intelligence and moral decisions. *Behavioral and Brain Sciences*, 28(4), 548-549.  
doi:10.1017/S0140525X05280093
- Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it? Reason in moral judgment. *Philosophical Explorations*, 9(1), 83-98.  
doi:10.1080/13869790500492680
- Flanagan, O. (2014). *Moral Sprouts and Natural Teleologies: 21<sup>st</sup> Century Moral Psychology Meets Classical Chinese Philosophy*, Milwaukee: Marquette University Press.

- Funk, C. M., & Gazzaniga, M. S. (2009). The functional brain architecture of human morality. *Current Opinion in Neurobiology*, 19(6), 678-681.  
doi:10.1016/j.conb.2009.09.011
- Gray, K., & Wegner, D. M. (2011). Dimensions of moral emotions. *Emotion Review*, 3(3), 258-260. doi:10.1177/1754073911402388
- Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2), 101-124.
- Greene, J. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 3*. MIT Press.
- Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322-323.  
doi:10.1016/j.tics.2007.06.004
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108. doi:10.1126/science.1062872
- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in cognitive sciences*, 6(12), 517-523
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814-834. doi:10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316(5827), 998-1002.

doi: 10.1126/science. 1137651

Haidt, J. (2012). Taste Buds of the Righteous Mind. In *The righteous mind: Why good people are divided by politics and religion* (pp. 112-127). Vintage.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*, 191-221.

Huebner, B., Dwyer, S., & Hauser, M. (2008;2009;). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, 13(1), 1-6.

doi:10.1016/j.tics.2008.09.006

Inwagen, P. v. (1997). Materialism and the psychological-continuity account of personal identity. *Philosophical Perspectives*, 11, 305-319.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & Language*, 25(5), 561-582. doi:10.1111/j.1468-

0017.2010.01401.x

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*, 7(4), 393-402. doi:10.1093/scan/nsr005

Kekes, J. (1984). 'Ought Implies Can' and Two Kinds of Morality. *The Philosophical Quarterly* (1950-), 34(137), 459-467. doi:10.2307/2219064

Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect, and the moral/conventional distinction. *Mind and Language*, 22(2), 117-131



- Koenigs, M., Young, L., Cushman, F., Damasio, A., Adolphs, R., Tranel, D., & Hauser, M. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908-911. doi:10.1038/nature05631
- Lewis, D. (1983). *Philosophical papers volume I*. New York: Oxford University Press. doi:10.1093/0195032047.001.0001
- McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, 45(3), 577-580. doi:10.1016/j.jesp.2009.01.002
- McMahan, J. (2013). Moral intuition. (pp. 103-120). Oxford, UK: Blackwell Publishing Ltd. doi:10.1111/b.9780631201199.1999.00007.x
- Mizrahi, Moti (2015). Ought, Can, and Presupposition: An Experimental Study. *Method 4* (6):232-243
- Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8), 319-321. doi:10.1016/j.tics.2007.06.001
- Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *Neuroimage*, 16(3), 696-703. doi:10.1006/nimg.2002.1118
- Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral

sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, 22(7), 2730-2736.

doi:10.1523/JNEUROSCI.22-07-02730.2002

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, 41(4), 663-685. doi:10.1111/j.1468-0068.2007.00666.x

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100(3), 530-542. doi:10.1016/j.cognition.2005.07.005

Pascual, L., Rodrigues, P., & Gallardo-Pujol, D. (2013). How does morality work in the brain? A functional and structural perspective of moral behavior. *Frontiers in Integrative Neuroscience*, 7, 65. doi:10.3389/fnint.2013.00065

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2(3), 511-527. doi:10.1111/j.1756-8765.2010.01096.x

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1), 163-177. doi:10.1111/j.1551-6709.2011.01210.x

Pujol, J., Batalla, I., Contreras-Rodríguez, O., Harrison, B. J., Pera, V., Hernández-Ribas, R., . . . Cardoner, N. (2012). Breakdown in the brain network subserving moral judgment in criminal psychopathy. *Social Cognitive and Affective Neuroscience*, 7(8), 917-923. doi:10.1093/scan/nsr075

- Pust, Joel, "Intuition", *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2019/entries/intuition/>.
- Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics*, 124(4), 813-859. doi:10.1086/675876
- Railton, P. (2017). Moral learning: Why learning? Why moral? And why now?. *Cognition*. <http://dx.doi.org/10.1016/j.cognition.2016.08.015>
- Sauer, H. (2012). Educated intuitions. automaticity and rationality in moral judgement. *Philosophical Explorations*, 15(3), 255-275. doi:10.1080/13869795.2012.706822
- Schüller, S. (2016). The effects of 9/11 on attitudes toward immigration and the moderating role of education: Effects of 9/11 on attitudes toward immigration. *Kyklos*, 69(4), 604-632. doi:10.1111/kykl.12122
- Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 34(13), 4741-4749. doi:10.1523/JNEUROSCI.3390-13.2014
- Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, 9(3/4), 331-352. doi:10.1007/s10892-005-3508-y

- Sinnott-Armstrong, W. (2006). Moral intuitionism meets empirical psychology. In T. Horgan & M. Timmons, *Metaethics after Moore* (pp. 339-367). Oxford: Oxford University Press.
- Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions as heuristics. *The moral psychology handbook, 1*, 246-272.
- Slovic, P., & Västfjäll, D. (2010). Affect, moral intuition, and risk. *Psychological Inquiry, 21*(4), 387-398. doi:10.1080/1047840X.2010.521119
- Stocker, M. (1971). Ought'and 'can. *Australasian journal of philosophy, 49*(3), 303-316.
- Strohming, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition, 119*(2), 295-300. doi:10.1016/j.cognition.2010.12.012
- Tobia, K., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts? *Philosophical Psychology, 26*(5), 629-638. doi:10.1080/09515089.2012.696327
- Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476-477. doi:10.1111/j.1467-9280.2006.01731.x
- Wang, A. B. (2018, October 9). Taylor Swift's endorsement of Democrats is followed by a spike in voter registrations. *The Washington Post*, Retrieved from <https://www.washingtonpost.com/arts-entertainment/2018/10/09/taylor-swifts->

[endorsement-democrats-causes-spike-voter-registrations/?noredirect=on&utm\\_term=.fd664e0b2872](#)

Wegner, D. M., & Gray, K. (2016). *The mind club: Who thinks, what feels, and why it matters*. Viking Adult.

Young, L., & Dungan, J. (2012). Where in the brain is morality? everywhere and maybe nowhere. *Social Neuroscience*, 7(1), 1-10. doi:10.1080/17470919.2011.569146

Young, L., & Koenigs, M. (2007). Investigating emotion in moral cognition: a review of evidence from functional neuroimaging and neuropsychology. *British medical bulletin*, 84(1), 69-79.