UNIVERSITÀ DEGLI STUDI DI MILANO

UNIVERSITÀ DEGLI STUDI DI MILANO

PhD program
Human Mind and its Explanations: language, brain and reasoning (38°)
Department of Philosophy "Piero Martinetti"

PhD Thesis

# CONTROLLED AND AUTOMATIC PROCESSES IN MORAL DECISION-MAKING

Renato Raia
R14007
ORCID id.: 0009-0006-6130-9257

Supervisors:
Prof. John Michael (University of Milan)
Prof. Stephen Butterfill (University of Warwick)

Head of the PhD program:
Prof. Andrea Sereni

2024-2025

*Morality is not something that should be written in philosophy books. It should be lived. Thus, this thesis is dedicated to the friends of my early adulthood in Rome (especially those who know nothing about philosophy) and the friends I made during my PhD in Milan and England. In good times and bad ones, they taught me morality and the joy of life.*

# Summary

## ABSTRACT IN ENGLISH

*A popular dual-process theory of moral decision-making, that put forward by Joshua Greene, proposes that deontological judgments are produced by affective intuitions while consequentialist judgements usually are the product of cognitive deliberation. Greene and colleagues (2001, 2004) have backed up this claim with empirical evidence coming from fMRI studies, cognitive load studies, and pathology. Greene (2008, 2014, 2017) has used this evidence to propose two normative derivations, the direct and indirect route, about deontological moral judgments. The arguments aim to show that, given their origin in affective and intuitive processes, deontological judgments led us astray in many modern moral problems, and in many philosophical problems like the famous trolley problem or Peter Singer's (1972) problem on the moral effect of spatial distance. In this thesis I attempt to flesh out a criticism against some aspects of Greene's dual-process model and to propose a new theory, by disentangling some theoretical classifications (intuition-deliberation, emotion-cognition, deontology-consequentialism) that, in Greene's theory, are mapped one upon the other. I argue that the emotion-cognition distinction is not a useful theoretical tool; and I follow the theories of the education of moral intuitions proposed by Jillian Craigie (2011), Hanno Sauer (2012a), Richmond Campbell and Viktor Kumar (2012) in positing a diachronic effect of deliberation over intuitions. I propose that the education of intuitions in the moral sphere is best characterized as a process of skill acquisition rather than habitualization, as this better accounts for its automatic yet rational nature (Christensen, Sutton and McIlwain, 2016; Stichter, 2018). Following an intuition by Peter Railton (2014), I provide a defense of deontological judgments from the normative claims of Greene's theory, and I offer a more balanced view of the role of intuition within moral decision-making, providing a reinterpretation of the role it plays in Haidt's Twin Problem, in the trolley dilemma, and in Singer's Puzzle. My reinterpretation is centered on the capacity of intuition to react to the general story proposed by these problems rather than on specific factors. Thus, I argue that intuition is much more nuanced and indeed necessary in morality. My own positive theory is centered around the notion that moral judgments might proceed from self-centered thoughts and feelings or from other-centered ones. I propose that broadly agreed upon judgments are those which find a balance between these two extremes, while deliberation is needed to reach more extreme and less common judgments. Finally, following a model by Kahane and colleagues (2018), I advance the hypothesis that consequentialist judgments might be perceived as more uncommon than deontological ones.*

## ABSTRACT IN ITALIANO

*Una popolare teoria a due processi del giudizio morale avanzata da Joshua Greene, propone che i giudizi deontologici siano prodotti da intuizioni affettive mentre i giudizi utilitaristi siano usualmente il prodotto di una deliberazione di tipo cognitivo. Greene e colleghi (2001, 2004) hanno supportato questa affermazione con evidenza empirica proveniente da studi di risonanza magnetica funzionale, da studi con carico cognitivo, e dalla patologia. Greene (2008, 2014, 2017) ha impiegato tale evidenza per proporre due derivazioni normative, la via diretta e la via indiretta, riguardo i giudizi morali deontologici. Tali ragionamenti mirano a dimostrare che, data la loro origine in processi affettivi ed intuitivi, i giudizi deontologici conducano a conclusioni errate in molti problemi morali moderni, e in molti problemi filosofici come il celebre dilemma del vagone o il problema sull'effetto morale della distanza spaziale proposto da Peter Singer (1972). In questa tesi, provo a sostanziare una critica contro alcuni aspetti del modello a due sistemi di Greene e a proporre una nuova teoria, districando alcune classificazioni teoriche (intuizione vs. deliberazione, emozione vs. cognizione, deontologia vs. utilitarismo) che, nella teoria di Greene, sono sovrapposte l'una all'altra. Argomento che la distinzione tra emozione e cognizione non è un utile strumento concettuale, e seguo le teorie sull'istruzione delle intuizioni morali proposte da Jillian Craigie (2011), Hanno Sauer (2012a), Richmond Campbell e Viktor Kumar (2012) nell'ipotizzare un effetto diacronico della deliberazione sulle intuizioni. Propongo inoltre che l'istruzione delle intuizioni nella sfera morale sia caratterizzata al meglio non come risultato di un processo di abituazione ma come l'acquisizione di una abilità complessa, perché ciò rende meglio conto della sua natura automatica e al tempo stesso razionale (Christiansen, Sutton, e McIlwain, 2016; Stichter, 2018). Seguendo un'intuizione di Peter Railton (2014), fornisco una difesa dei giudizi deontologici dalle affermazioni normative della teoria di Greene, e offro una caratterizzazione più bilanciata del ruolo dell'intuizione all'interno del giudizio morale, fornendo una reinterpretazione del ruolo che essa gioca nel problema dei gemelli di Haidt, nel dilemma del carrello, e nel problema della distanza proposto da Singer. La mia reinterpretazione si concentra su fatto che l'intuizione, piuttosto che concentrarsi su specifici fattori dei problemi trattati, reagisce alla narrazione generale in essi contenuta. Dunque, argomento che l'intuizione è molto più sofisticata e senza dubbio necessaria nel giudizio morale. La mia proposta in positivo si concentra sull'idea che i giudizi morali possono procedere da pensieri e sentimenti ego-riferiti o etero-riferiti. Io propongo che i giudizi morali che trovano ampia accettazione siano quelli che si trovano in un bilanciamento tra questi due estremi, mentre la deliberazione è necessaria per raggiungere giudizi più estremi e meno comuni. Infine, seguendo un modello proposto da Kahane e colleghi (2018), avanzo l'ipotesi che i giudizi utilitaristi possano essere percepiti come meno comuni di quelli deontologici.*

# INTRODUCTION

While the moral prescriptions of many religions and philosophical systems aspire to have a universal scope, the lay morality of most people shows strong personal biases: in our moral judgments, we show a preference for our family members, friends, and compatriots, and we abstain from acting in ways which would nonetheless produce the greater good if this means inflicting harm on innocents. Against these biases, the philosophical theory of classical utilitarianism (also known as *consequentialism*), developed in the 18th century by philosopher Jeremy Bentham and famously defended by John Stuart Mill, holds that we should always act in ways that maximize aggregate well-being. As they ask us to extend our circle of moral concern to all of humanity (or even all sentient beings) without showing preference for those close to us and as they prescribe us to act in favor of the greater good even if this means performing harm on specific innocents, the prescriptions of utilitarianism are at odds with many widespread moral beliefs, and have found fierce resistance from both lay people and many moral philosophers. However, the fact that consequentialist judgments are often against the common sentiment has been used not only against consequentialism but also in favor of it. The defenders of consequentialism have argued that our *moral intuition* is the source of negative biases and represents an undue influence of our primitive emotional systems on the rational conclusions of consequentialist calculation. They have argued, in short, that *deontological* judgments (those preventing us from doing harm even if greater good would come out, and, more generally, those deriving from the considerations of factors besides the maximization of welfare) are particularly related to an *automatic*, and often *emotional*, kind of psychological processing; while consequentialist judgments are particularly related to a *deliberative* kind of psychological processing. For example, the theory proposed by Joshua Greene can be articulated into two components: first, it is shown, on the basis of a significant wealth of empirical evidence, that deontological judgments are related to intuition and emotion, while consequentialist judgments are related to deliberation and cognition; second, a reasoning is made to show how from the first claim comes the conclusion that consequentialist judgments ought to be preferred to deontological ones. This thesis proposes some criticism on the traditional dual-process theory of moral judgment as offered by Greene at tries to envision a new positive model.

Here is a chapter-by-chapter summary of the thesis, hopefully helping the reader to orient himself within the main argument:

CHAPTER 1 – THE DUAL-PROCESS THEORY OF MORAL DECISION-MAKING

The aim of Chapter 1 is to introduce Greene's dual-process theory by comparing it with another dual-process theory of moral judgment, namely Jonathan Haidt and colleagues' *social intuitionism*, which represents a modern version of sentimentalism. Both Haidt's theory and Greene's theory can be seen as reactions to the intuitionist challenge against moral rationalism, and both agree on the idea that the emotion-cognition distinction can be mapped upon the intuition-deliberation distinction. However, contrary to Haidt's sentimentalism, Greene's dual-process model has a *rationalist* undertone, which is summarized in the *Central Tension Principle*. According to the Central Tension Principle, deontological judgments are mostly the product of affective intuitions, while consequentialist judgments are mostly the product of cognitive deliberation. So, in this principle, the three distinctions between deontology and consequentialism, emotion and cognition, and intuition and deliberation are mapped one upon the other.

After having discussed basic vocabulary in the second section and having introduced the Central Tension Principle in the third, in the sections four and five I will present the main empirical evidence in favor of the principle. Such review does not aspire to be exhaustive, but just to show how some of the evidence (namely, that related to reaction time and cognitive load) is at best unconclusive, while further evidence (namely, that coming from fMRI studies) only lends support to the dual-process theory if one assumes a particular view of the relation between psychological function and brain area.

CHAPTER 2 – DEBUNKING ARGUMENTS AGAINST DEONTOLOGICAL JUDGMENTS

Chapter 2 is devoted to the normative conclusions Greene tries to deduce from empirical evidence, focusing on the main argumentative line which supports them. The first three sections are dedicated to the so-called "direct route" argument. Two examples of this argument are made: one related to the trolley problem and the other related to Singer's Puzzle. In these problems, two scenarios are compared: in one we tend to judge deontologically, while in the other we tend to judge in a consequentialist manner. Greene hypothesizes that the reason for this difference in judgement is that the former scenarios contain factors, like *personal violence* (in the trolley dilemma) or *spatial proximity* (in Singer's Puzzle), that evoke our emotions, steering our judgment into a deontological direction. Greene (2008) then hypothesizes that the reason why personal violence and spatial

proximity evoke stronger emotional reactions in us is because they are related to *ancestral needs* or conditions of our species, like the need to control in-group violence or the impossibility to know about far away people. Deontological judging is thus related to emotions which react to factors that are *morally irrelevant*, and that we only care about given the peculiarities of our phylogenesis. Then Greene argues that deontological philosophy is a *rationalization* of these emotional reactions and that thus we ought to prefer consequentialist philosophy.

The fourth and fifth sections are instead devoted to the "indirect route" argument which can be seen as an expansion and a generalization of the direct route. Moral intuitions and emotions do not come only as a result of evolutionary experience, but also as a product of cultural and personal experience. However, *the way in which intuitions are formed is rigid*. Here, a comparison is made between Greene's indirect route and the dual-process theory by Fiery Cushman (2013), which also proposes that deontological judgments are related to quite rigid learning algorithms, while consequentialist judgments employ costlier but more flexible algorithms. The rigidity of the processes behind the formation of intuition means that it, and the related deontological judgments, are set to yield wrong conclusions in moral problems related to the modern world and philosophical moral problems where the relationship between causes and effects is unusual.

The next four chapters can be seen as a reaction to the arguments of Greene's dual-process theory and constitute the positive proposal of the thesis.

CHAPTER 3 – BEYOND THE EMOTION-COGNITION DIVIDE

The critical target of Chapter 3 is what Saunders (2016) calls the "standard approach" to moral decision-making which consists in the view that emotion and cognition are two encapsulated faculties which can only have input-output interactions. In this chapter, I make the general claim that t *the emotion-cognition divide is not a good framework for interpreting moral judgment*. After having introduced the standard approach in the first section, in the second section I employ the example of moods as non-intuitive emotional states which affect moral decision-making, and thus as a way to show how *the emotion-cognition divide cannot be easily mapped onto the intuition-deliberation distinction*. The second part of Chapter 3 can be seen as a reaction to the empirical evidence collected in favor of the Central Tension Principle, which focuses on one region in particular: the amygdala. Greene argued that the amygdala is related to emotion and that the activation of the amygdala during the production of deontological judgments shows how these judgments trace their origin in affective processes. In the two central sections of the chapter, I present the interpretation of

the amygdala as the locus of implementation of a modular emotion in particular (fear) and I show the evidence we have for the conclusion that this interpretation is not correct. I make a comparison between the reinterpretation of the role of the amygdala and the reinterpretation of the role of Broca's area. We previously believed that Broca's area was devoted to the processing of linguistic structures, and, in particular, of language's syntax. Then, we discovered that Broca's area is also sensitive to other kinds of syntactic structures, even when expressed in a non-linguistic modality. The earlier interpretation of Broca's area as a language-specific area was thus put in doubt by the discovery that it reacted to linguistic sequences *not by virtue of them being linguistic*, but just by virtue of them being syntactic structures. Similarly, the evidence recently collected about the amygdala links the area with functions related to *attention* and attribution of value to stimuli. Threatening stimuli particularly activate the amygdala because they are stimuli that elicit attention. The early interpretation of the amygdala as the seat of the modular emotion of fear is thus disputed by the evidence that it might react to threatening stimuli *not in virtue of them being threatening*, but just as a consequence of the fact that they elicit attention. It is important to note that this new interpretation ascribes at the amygdala functions which put the region at the crossroad between cognition and emotion, making it impossible to consider it a purely affective area.

The last section of Chapter 2 can be seen as a generalization of the conclusions reached about the amygdala. Evidence about the reuse of motor, perceptual, and affective areas of the brain during the performance of tasks related to conceptual function shows how emotion and cognition cannot be localized in the brain, thus disputing the standard approach idea that these are encapsulated faculties only having input-output interactions.

CHAPTER 4 – THE EDUCATION OF INTUITIONS

The main claim of Chapter 3 is that *moral intuitions are not an encapsulated process*, but they are influenced by deliberate processing, giving rise to a form of educated intuitions.

The aim of chapter three is to present the theories of the education of intuitions proposed by Hanno Sauer (2012a), Jillian Craigie (2011), Richmond Campbell and Viktor Kumar (2012), and Michelle Maiese (2014). These theories have different aims: in particular, Sauer's goal is to defend a form of moderate rationalism from the criticism of sentimentalist theories like social intuitionism, while Craigie's goal is to propose a more integrated relation between affective and cognitive processes than that offered by Greene's model. While the aims differ, there are a lot of commonalities between the theories of education of intuitions presented in this chapter. All these theories propose that, while most moral judgements are experienced as fast and effortless, acts of private reflection still exert a strong influence over moral judgments, as reasoning on one's previous intuitions makes future intuitions

sensitive to previously overlooked factors. The intuitive judgments of a human adult require appreciation of many complex concepts. The education of intuitions can be thus framed as the acquisition by intuition of responsiveness to complex, "deliberative" factors, or as the migration of slow deliberative evaluations into a faster intuitive modality. The education of intuitions gives rise to a form of intuitive judgment which can be properly considered to be rational and deliberative. As such, theories of the education of intuitions overcome the standard approach to moral decision-making, by positing a direct influence of reasoning over the inner mechanisms of intuition.

Theories of the education of intuitions also make a hypothesis about the reason why such influence of deliberation over moral intuitions was not captured by the early dual-process theories of moral judgment. The experiments performed to test these theories focused a *narrow timespan* of moral judgment, going to look at the psychological processes which immediately preceded the judgement, processes which are almost always intuitive. If we adopt instead a *diachronic perspective*, we will also consider the factors that influence our intuitions, shaping them over the course of our life as moral subjects. Here, we might find the influence of deliberation.

## CHAPTER 5 – DEFENDING MORAL INTUITIONS

Chapter 4 constitutes an attempt to defend the validity of deontological judgments from the arguments of both the direct route and the indirect route. According to Greene's theory, deontological judgments are particularly related to intuition. Here, I will not dispute this claim, but I will argue that moral intuition should be revalued, and that the fact that a judgment is intuitive does not constitute a reason to doubt its warrant. The reason why Greene thinks so is that he argues, in the direct route, that intuition is responsive to specific factors which were only relevant in our ancestral past. I will argue instead that *moral intuitions (and deontological judgments) do not respond to morally irrelevant factors and, more generally, they do not respond to any easy-to-pinpoint factors of moral dilemmas.* All the preceding considerations about the education of intuitions open the possibility that intuition might not be activated by the presence of simple factors within moral problems such as Haidt's twin problem, the trolley problem, or Singer's Puzzle. This is the main idea behind my defense of moral intuition and deontological judgment and my reinterpretation of the results coming from these mental experiments. I will propose that moral intuitions react to the *general narratives* inherent in these problems.

Rather than being elicited, via disgust, by the simple presence of incest in Haidt's twin problem, intuition might well consider the risks the twins incurred in which, while not resulting in a direct damage to them, still allow for the twins to be rightfully condemned for

their foolishness. Similar considerations also apply to the trolley problem: rather than reacting to the simple presence of personal violence, intuition might consider all a series of other aspects of the problems. I propose that it is only because the scenario is a mental experiment, and all other considerations are deemed irrelevant, that intuition is forced to focus on the personal violence exerted in one of the cases. Finally, we have evidence that spatial distance is confounded with all a series of other variables and is not what our intuition is reacting to in Singer's problem.

As for the indirect route, one of the major problems I identify is that in the problems Greene thinks only deliberation can properly tackle one still needs some kind of deontological evaluation before the consequentialist harm calculation can start. Indeed, deontological judgments are what lead us into different conclusions about these moral problems, and this seems not a feature that deliberation can really help to solve.

CHAPTER 6 – THE ATTUNEMENT MODEL OF MORAL DECISION-MAKING

Chapter 6 contains my positive proposal beyond the standard dual-process approach to moral cognition.

In the first section of Chapter 6, I try to find possible ways to improve theories of the education of intuitions. There I look at some potential limitations of the theories of education of intuition as presented for example by Sauer. One limitation is that Sauer characterizes the process of education in terms of *habitualization*, but habits are usually described as quite recalcitrant to rational influence. Furthermore, Sauer's theory leaves unspecified if there are forms of overt moral reasoning which cannot be automated into an intuitive format. Finally, Sauer seems to think that the rationality of educated intuitions is dependent on the ability of deliberation to recover and verbalize the true reasons behind a judgment, but this requirement seems too strict. I propose that a possible way to overcome these limitations is to conceive of the process of the education of intuitions not in terms of habitualization but as a process of *skill acquisition*. A specific interpretation of skill acquisition, that advanced among others by Christensen, Sutton, and McIlwain (2016) and Stichter (2018), is particularly apt to tackle the limitations of traditional theories of the education of intuitions. According to this interpretation, we have evidence to think about skill acquisition as a process which involves both deliberative and intuitive aspects, and, in particular, as a process where the control of low-level aspects of the implementation of action is automated and made intuitive while higher-order strategic control remains under deliberate control. Furthermore, the expert in a skill needs not to be able to fully verbalize her competence.

According to the Central Tension Principle, the distinctions between deontology and consequentialism, cognition and emotion, and intuition and deliberation can be mapped on

one another. In Chapter 2, we will see how the divide between emotion and cognition cannot be meshed with that between intuition and deliberation, and how, more generally, wit is a bad framework by which to interpret moral decision-making. In Chapter 3, we will see how the distinction between intuition and deliberation is less strict than previously thought. In Chapter 4, we will look at how that distinction cannot even be founded upon the idea that intuition responds to morally irrelevant factors. Chapter 5 continues this work of conceptual de-confounding as, following the research by Guy Kahane and colleagues (2012), I will show how even the distinction between deontology and consequentialism is an inappropriate lens through which to look at the intuition-deliberation divide. I will in fact argue for the claim that *consequentialist judgments are not related to greater deliberative capacity*. This leads to a "banal" dual-process model, in which those judgments that usually find support by people within a culture (which I call *common* moral judgments) are related to intuition while more controversial moral judgments (*uncommon* ones) require deliberation to be reached.

Chapter 5 then continues by offering a positive model of the distinction between intuition and deliberation in moral judgment. I believe there is a difference in the role played by intuition and deliberation in moral judgment, but this difference is not properly captured by the conceptual grid of the Central Tension Principle. I will not offer a general view of the relation between intuition and deliberation in moral judgment, that would be a really difficult task because intuitive psychological processes and deliberate psychological processes aren't internally cohesive: there are many kinds of intuition and many kinds of reflection, and, as the education of intuition theories suggest, these processes might be much more deeply interconnected than previously believed.

To clarify, I will introduce a further set of evidence which is usually interpreted to favor the Central Tension Principle, that coming from pathology. I will focus on the ventromedial prefrontal cortex (vmPFC), which Greene interprets as an affective region of the brain, given that damage to it leads to a compromised emotional life. However, while it is true that damage to the area also leads to an increase in the frequency of consequentialist judgments when the subject is confronted with the trolley problem or similar other scenarios, when confronted with other problems of choice, like the Ultimatum Game, subjects with a damage to the area make choice related to an increased emotional sensitivity rather than stunned emotions. This leads to the possibility that, rather than being related with emotion specifically, the ventromedial prefrontal cortex is connected with processes of *attunement* to the necessities of social life. This attunement is characterized by an inextricable union of both cognitive and affective aspects and damage to the area leads to a misalignment between the subject's thoughts and emotions and the requirements of social life, with more self-centered emotions and thoughts.

My positive model begins with the idea that the distinction between being self-centered and being other-centered (caring of other people and being sensitive to their expectations) is a conceptual distinction with real correlates. I hypothesize that moral common judgments, those usually supported by intuition, are those that derive from a culturally mediated balance between self-centeredness and other-centeredness, while uncommon moral judgments are those that are found at the extremes of the scale. Pathology can make uncommon moral judgments intuitive, by making the subject, for example, more self-centered in both his feelings and his reasoning. Following an intuition by Kahane and colleagues (2018), I propose that consequentialist judgments often come out as uncommon, because they derive from excessive self-centeredness (negative consequentialism, that is consequentialism as acceptance of instrumental harm) or because they prescribe excessive other-centeredness (positive consequentialism, that is consequentialism as universal beneficence). In the last section of Chapter 6, I make some further considerations on the model I propose, hinting at possible connections between it and other relevant theories in moral cognition including moral disengagement theory, Moral Foundations Theory and Shaun Nichols' theory of moral learning.

My gratitude goes to my supervisors John Michael and Stephen Butterfill for their intellectual and personal support. I am also grateful to the suggestions of the external examiners András Szigeti and Hanno Sauer. I also want to thank my parents, who supported me in my choice to study philosophy despite the perils inherent in such choice. Finally, a special and very personal mention goes to my friends Antonio D'Aiello and Andrea Donato.

# 1. THE DUAL-PROCESS THEORY OF MORAL DECISION-MAKING

> "Nothing is more usual in philosophy, and even in common life, than to talk of the combat of passion and reason."
>
> - *David Hume, "A Treatise of Human Nature" (1740)*

According to the dual-process paradigm of human higher psychological functions, they are the result of the interplay between an automatic kind of processing and a controlled one (Evans, 2008; Frankish and Evans, 2009; Kahneman, 2011). Precisely characterizing the distinction between the two types of processing has proven a controversial task because, over the course of the decades, different authors have ascribed to the types of processing properties which are often contradictory or founded on ambiguous dichotomies, and which do not perfectly align (Evans and Stanovich, 2013). I will use the term "intuition" and "intuitive" to refer to psychological activity of the automatic type, and the term "deliberation", "deliberative", "reasoning" and "reasoned" to refer to psychological activity of the controlled type. The purpose of this section is to introduce Joshua Greene's dual-process theory of moral decision-making and compare it with Jonathan Haidt's social intuitionist model. Both these theories can be seen as a reaction to the evidence we have about automaticity and confabulation in human moral behavior, evidence which makes us doubt deliberate control over our judgments and actions. The conclusions Haidt and Greene draw from this evidence are markedly different: Haidt accepts that human moral behavior is almost entirely dependent on intuition, and that, at least at an intrapersonal level, the only purpose of deliberation is to confabulate explanations for the decisions taken by intuition; while Greene believes that deliberation can arrive at a further set of moral conclusions which are exclusive to it.

## 1.1. SOCIAL INTUITIONISM

The starting point of the dual-process theories of moral decision-making is the *intuitionist challenge* against moral rationalism. While early theories of moral decision-making emphasized the role of conscious deliberation in moral judgment and the rational nature of the moral agent, rationalism has been more recently challenged by findings about the pervasiveness of *automaticity* in human moral behavior and of *confabulation* of the explanations we make of such behavior: for example, Haidt, Björklund and Murphy (2000) administered little stories like the following to subjects, asking them if the action performed by the protagonists was morally acceptable:

TWINS DILEMMA

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

While most subject found the act of incest morally inappropriate and initially produced arguments about it, the story is told in such a way to prevent most reasonable objections, and when subjects became aware of this, they often found themselves into a state of *moral dumfounding*, insisting that the act was wrong but being unable to put forward reasons to defend their judgment. The authors concluded that the real reason why the subjects hold the incest to be wrong was their affective intuition (in this case, disgust), and that the previous arguments were just a confabulation to justify their emotion.

Evidence about automaticity and confabulation has been used to argue in favor of *intuitionist* theories, which highlight the fast, automatic, and unconscious nature of moral decision-making. Intuitionist theories tend to be *sentimentalist* in nature, claiming that morally relevant intuitions have an essentially affective character. For example, Haidt (2001, p. 818) 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." This definition emphasizes the lack of conscious awareness and an affective valence as the main properties of moral intuitions. Similarly, Greene (2008, p. 41) writes that: "[…] while the term 'emotion' can refer to stable states such as moods, here we will primarily be concerned with emotions subserved by processes that, in addition to being valenced, are quick and automatic, though not necessarily conscious." It should be stressed that the concept of emotion, with all its conceptual fuzziness, is distinct from the concept of intuition, and the same holds true for the pair deliberation-cognition. Stating that most intuitions which are relevant in moral judgment have an affective component, and on the other side that moral deliberation is a "colder" process is by itself a discovery, and something that should be defended with evidence and argument rather than just assumed (more on a characterization of emotion and cognition in the second section of this chapter).

One intuitionist theory systematizing the consequences for rationalism of the evidence about automaticity and confabulation is *social intuitionism*, proposed by Haidt and

colleagues (Haidt, 2001; Haidt and Björklund, 2008). In a nutshell, social intuitionism holds that the true proximal causal origin of most moral judgments is to be found in *affective intuitions*, and that conscious processes of deliberation serve the social function of producing justifications for judgments that have already been made by intuition. Social intuitionism is articulated into six links, each reflecting a psychological process connecting intuition, judgment, or deliberation in the moral sphere. Together the links constitute a positive proposal about the relation between moral affective intuitions and moral decision-making, and a critical take on moral rationalism. The bulk of social intuitionism is contained in the first two links which state that moral judgments are the product of moral intuitions and that moral deliberation, defined as conscious mental activity consisting in the transformation of information in order to reach a moral judgment, is not the causal origin of the judgments: the subject deliberates *after* a judgment has already been made, and deliberation serves the purpose of searching for arguments that defend the judgment (Haidt, 2001, p. 818). The definition of intuition and deliberation in social intuitionism make them *mutually exclusive*: either a judgment is the product of intuitions, or it is the product of a deliberate process of inferential reasoning. Social intuitionism claims that we have evidence that moral judgment is intuitive and that moral deliberation is post-hoc and that from these two facts we can deduce that moral deliberation is causally ineffective confabulation.

While the first two links of social intuitionism express the "intuitionist" part of the model, the further two contain its "social" part and explain why nature has endowed us with deliberation even though it is causally powerless (Haidt, 2001, p. 818-819; Haidt and Björklund, 2008, p. 190-192). We often reason after a judgment when we feel the pressure to justify our judgments in front of others. While deliberation can be interiorized by the modern subject, its primary function is eminently social: it serves the purpose of justifying to others our judgments and of persuading others to adopt them. Acts of deliberation can induce the adoption of new moral judgments into others, mostly by generating newly affective intuitions into them. What is denied by social intuitionism is the possibility of a *direct* effect of *intrapersonal* moral reasoning over one's moral convictions, at least in normal circumstances. Haidt writes that (2001, p. 819): "The core of the [social intuitionist] model gives moral reasoning a causal role in moral judgment but only when reasoning runs through other people." The judgments made by one person can also influence judgments made by another even without passing through reasoned deliberation (Haidt, 2001, p. 819): as humans are the only animal who lives in large groups constituted by non-kin, they have peculiar psychological adaptations making them sensitive to what other people think and feel. People tend to conform to the judgments of friend and allies and our moral judgments "are strongly shaped by what others in our 'parish' believe, even when they don't give us

any reasons for their beliefs" (Haidt and Björklund, 2008, p. 193). With these two further claims of social intuitionism, deliberation re-acquires causal effectiveness, albeit *interpersonally*, as it helps to convince others of one's moral standing. This effect is *indirect*: I can use my deliberation to induce new judgments into others, mostly by generating new affectively valenced intuitions in them. On the contrary, rationalists think that *private* moral reasoning can have *direct* effects on moral judgments: that my acts of reasoning can induce a change in my own moral judgments.

The final two links of social intuitionism are "special" in that they model exceptions to the average course of things and can be seen as concessions to rationalism. Despite the evidence collected by Haidt and others, it happens sometimes that people engage in genuine non-post-hoc private moral reasoning. People can arrive at a judgment by sheer force of logic, overriding their initial intuition or they may activate new intuitions by thinking about a situation (Haidt, 2001, p. 819; Haidt and Björklund, 2008, p. 193-196). It should be stressed that Haidt and colleagues believe that these processes occur very rarely: for private deliberation to overcome the effect of affectively valenced intuitions over judgments, the initial intuitions have to be weak and the processing capacity of the subject has to be high, and the subject has to have some special need to trust his reasoning more than his intuition. I think that these concessions do not actually make social intuitionism compatible with rationalism, insofar as rationalists believe that the effects of private reasoning over one's moral judgments are systematic and somewhat frequent.

To summarize, social intuitionism is a dual-process theory of moral decision-making that employs the findings about automaticity in moral judgment and confabulation in moral deliberation to argue that deliberation is not causally effective in moral judgment, thus taking a critical stance against rationalism. Social intuitionists hold that the true cause of moral judgments at an intrapersonal level are quick flashes of emotional appraisal, outside of the scope of rational control. Reasoning is causally powerless, and it serves the function of socially justifying our judgments by producing arguments whose conclusion was set in advance by intuition. Social intuitionists allow for reasoning to play a role in influencing moral judgments, but only interpersonally and indirectly; holding that whatever effect intrapersonal moral reasoning play is highly rare and unsystematic.

### 1.2. BASIC TERMS

While Haidt and Greene agree that moral intuitions are affectively valenced, they make different diagnoses of the consequences for rationalism of the evidence about automaticity and confabulation, with Haidt's social intuitionist model being sentimentalist in spirit, and Greene's dual-process theory having a rationalist normative conclusion. While social intuitionism envisions a division of labor between intuition and deliberation

concerning the different *kinds of role* they play in moral judgment, with intuitions being the true cause of all judgments while moral deliberation is mostly confabulation, the division of labor envisioned by Greene's theory can be found in the different *kinds of judgment* intuition and deliberation produce.

According to Greene (2008, p. 41), rationalism, as exemplified by Kohlberg's (1971) theory, and sentimentalism, as exemplified by Haidt's (2001) model, are two "extreme" position, because they consider *all* moral judgments to be either as the product of intuition or deliberation. These theories are thus insensitive to the classical distinctions within normative ethics (deontology, consequentialism, virtue ethics et cetera). Greene (2008, p. 36) writes: "his [Haidt's] radical thesis is intended, if only implicitly, to apply equally to the adherents of all moral philosophies." The traditional view within philosophy holds instead that consequentialist considerations are linked with emotional intuitions, while deontological judgments with metaethical rationalism. Contrary to the "extreme" views, in traditional philosophy the distinction between consequentialism and deontology makes a metaethical difference. From the historical view, Greene derives the idea that deontological judgments and consequentialist ones represent two subclasses of moral verdict with different psychological origins.

The first relevant distinction to understand contemporary research in moral cognition is thus that between deliberation and intuitions, with moral intuitions also characterized as affective ones. Moral intuitions tend to employ *emotional representations*, while deliberation tends to employ *cognitive representations*. But what is the difference between these kinds of representations? Emotional representations are "behaviorally valenced" (Greene, 2008, p. 40) and that have "direct motivational force" (Greene et al., 2004, p. 397); while cognitive representations are "inherently neutral" and "do not automatically trigger particular behavioral responses or dispositions" (Greene, 2008, p. 40). According to Greene, the distinction between affective processes and cognitive ones, and more generally the distinction between intuition and deliberation, exemplifies a solution to a problem of the general design of psychological functions: the "trade-off between *efficiency* and *flexibility*" (Greene 2014, p. 696, italics in the original). Intuitions are highly efficient, because they drive the organism experiencing them towards an adaptive pattern of behavior in an automatic and unconscious fashion, without the need for a costly and slow analysis of the situation, but their immediate motivational force makes them quite unflexible; while neutral representations, which do not immediately presuppose a particular behavioral response, are necessary for "reasoning, planning, manipulating information in working memory, controlling impulses and 'higher executive functions'" (Greene, 2008, p. 40). These abilities enable behaviors that serve longer-term goals and ensure a more flexible relation with one's

environment. While most animals only possess efficient but inflexible representations and processes (Greene, 2014, p. 698), some primates have also developed a more behaviorally dethatched and flexible system. And while humans undoubtably possess the most complex system for deliberation and cognition, their behavior is still largely guided by affect-laden intuitive reactions. Our psychology is, at many levels, the result of the interaction between efficient processes and flexible ones.

The trade-off between efficiency and flexibility is illustrated, for example, by the *now vs. later* tension (Greene, 2014, p. 697). When put in front of a valuable resource, like food or money, our efficient processes urge us to acquire it immediately. Our deliberative processes can instead prevent us from consuming it now in order to satisfy some longer-term goal (not eating cake now for a slimmer waistline in the future, not taking the money now to be awarded a larger sum later, and so on). But our deliberative processes, whose operations "are typically conscious, experienced as voluntary, and often experienced as effortful" (Greene, 2014, p. 697), have a limited capacity: putting pressure on them, by requiring the completion of some further tasks which engages them, reduces their overall efficacy. Indeed, cognitive load is one of the main methods by which dual-process theories of decision-making have been tested (Evans & Stanovich, 2013, p. 232). When some kind of cognitive load is administered, the resources of deliberative processes are depleted, and the subject will tend to consume a valuable resource now even if this conflicts with her long-term goals (Shiv & Fedorikhin, 2002).

In analyzing Greene's theory, as it holds true for most well-grounded philosophical theories, it is important to pay attention to technical vocabulary. In this thesis, as said above, I characterize intuition as a kind of psychological processing which is mostly fast, automatic, and unconscious. The opposite of an intuitive process is a *deliberate* process, so deliberation is the opposite of intuition. Deliberation is characterized by being slow, reflective, and mostly conscious. The intuition/deliberation distinction should not be meshed with the emotion/cognition distinction. What emotions themselves are is a complex philosophical and scientific question, but, in this thesis, I will accept Greene's suggestion and characterize emotional processing as a *behaviorally valanced* kind of psychological processing. The opposite of an emotional process is a *cognitive* process, so cognition is the opposite concept to emotion, and a cognitive evaluation of the situation allows us to assess a situation in a more neutral and dispassionate way. In this work, I will thus use the terms "cognitive" and "cognition" in their narrow sense of being opposed to emotion, rather than in the wide sense of them referring to all psychological or mental processes. Whether the emotion-cognition distinction can be mapped onto the intuition-deliberation distinction is an open question whose answer depends on the specific theory under consideration. Greene's answer, as we

saw, is mostly yes, as, at least insofar as moral judgment is concerned, intuitive processes tend to be emotional while deliberate processes tend to be cognitive, while my own answer, as we shall see, is mostly no.

We thus have a mapping between two kinds of moral judgments (deontological vs. consequentialist) and two kinds of psychological processes (affective intuition vs. cognitive deliberation). Greene (2008, p. 41) talks of "four basic empirical possibilities" that come up crossing the two different kinds of moral judgment with the two kinds of psychological processes. He says that: "First, it could be that both kinds of moral judgment are generally 'cognitive', as Kohlberg's theories suggest. At the other extreme, it could be that both kinds of moral judgment are primarily emotional, as Haidt's view suggests." From this, it can be concluded that, for Greene, Haidt's theory and Kohlberg's one are at the opposite side of the spectrum in answering to the question "are moral judgments emotional or deliberative?". Greene continues: "Then there is the historical stereotype, according to which consequentialism is more emotional (emerging from the 'sentimentalist' tradition of David Hume [1740/1978] and Adam Smith [1759/1976]) while deontology is more 'cognitive' (encompassing the Kantian 'rationalist' tradition [Kant, 1959])."

Mapping the distinction between consequentialism and deontology into a distinction between kinds of psychological processes is conceptually problematic. This is because deontology and consequentialism are usually taken to be moral philosophies, and thus approaches whose definition fully depends on the philosophers defining them. Deontology is defined as the approach putting emphasis on moral rules, articulated in terms of rights and duties; while consequentialism maintains that the moral value of an action is a function of its consequences alone (Greene, 2008, p. 37). A judgment resulting from, say, an emotional impulse would thus not count *by definition* as a deontological judgment, because it has not been made out of respect for moral rules. As the aim of his empirical hypothesis is to link deontology and consequentialism to specific psychological processes, Greene must redefine the concepts of deontology and consequentialism in a way that makes them tractable by means of empirical analysis. He writes (2008, p. 37-38, italics in the original):

> "[…] the terms 'deontology' and 'consequentialism' refer to *psychological natural kinds*. I believe that consequentialist and deontological views of philosophy are not so much philosophical inventions as they are philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking that have been part of the human repertoire for thousands of years. […] moral philosophies […] are just the explicit tips of large, mostly implicit, psychological icebergs."

According to Greene, deontology and consequentialism as psychological natural kinds can be defined from their *functional* properties. This is because deontologists and consequentialists will have a series of practical disagreements. A deontologist and a consequentialist will suggest different courses of action in the same moral dilemma. Take for example the famous *switch dilemma*. A runaway trolley is going to run over five people, which would die if the trolley was not diverted. The only way to save them is to hit a switch that will turn the trolley onto a different track, where there is one person which will die as a result of the switching. It is morally permissible to hit the switch and divert the trolley? According to Greene, this dilemma is an effective way of pitting deontological and consequentialist considerations against one another. A consequentialist would say that killing one person and saving five minimizes overall harm. He would thus suggest that it is permissible to hit the switch in consideration of the positive consequences that such action has. A deontologist would disagree, stating that producing a death by means of one's action is an inherently worse moral choice than letting people die as a consequence of one's inaction. In other words, the deontologist would argue that there is a stronger moral rule against killing than against letting die.

In reality, many *actual* deontologists have tried to justify hitting the switch. They tried to explain why hitting the switch in the switch dilemma is consistent with an overall deontological framework. But from the perspective of Greene's functional definition of deontology and consequentialism, this attempt is irrelevant. Finding it morally appropriate to hit the switch in the switch dilemma is a judgment "more easily justified in terms of the most basic consequentialist principles" (minimizing overall harm) and deontologists need "a lot of fancy philosophizing" to show that such judgment is also consistent with a deontological framework (Greene, 2008, p. 39). A judgment of this sort is thus a "characteristically consequentialist judgment" although is made by a philosopher who usually considers herself a deontologist. On the other hand, the judgment in favor of characteristically deontological conclusions (like the judgment that it is not morally appropriate to hit the switch) is a "characteristically deontological judgment", even if a consequentialist were to endorse it. I will not assess here the validity of this functional definition of deontology and consequentialism, as I am interested in the empirical hypothesis which requires such definition.

### 1.3. THE CENTRAL TENSION PRINCIPLE

We have a debate in moral psychology pitting rationalists against sentimentalists. In traditional philosophy instead, it is believed that classical distinctions in types of moral philosophies make a metaethical difference, with some kinds of moral judgments being more strongly related to certain kinds of psychological processes. More precisely,

deontology is related to deliberation, while consequentialism is more emotional. The dual-process theory of moral judgment offered by Greene and colleagues can be seen as a position of *synthesis* within the debate between cognitivists and sentimentalists. It proposes in fact that some kinds of judgments are produced by affective intuitions while other kinds of judgments are a result of cognitive deliberation. The dual-process theory can also be seen as an *inversion* of the traditional philosophical view. It holds in fact that deontological judgments are emotional and intuitive while consequentialism is cognitive and deliberative. According to the dual-process theory, the philosophical tension between schools of thought (deontology and consequentialism) is a manifestation of the psychological tension between efficiency (represented by processes which are most often affectively characterized) and flexibility (represented by processes which are not immediately behaviorally valenced). This is summarized by what Greene (2014, p. 699) calls the Central Tension Principle:

> CENTRAL TENSION PRINCIPLE: "Characteristically deontological judgments are preferentially supported by automatic emotional responses, while characteristically consequentialist judgments are preferentially supported by conscious reasoning and allied processes of cognitive control."

The Central Tension Principle is an empirical proposal about the functioning of moral decision-making, and it is arguably the main empirical claim offered by the dual-process theory of moral decision-making. The Central Tension principle serves as the empirical premise of two routes of normative derivation made by Greene to conclude that consequentialist philosophy is to be preferred to deontological philosophy. The arguments start from the assumption that deontological judgments depend on affective intuitions, then they aim to show that, given their origin, intuitions are unlikely to track the moral truth. A further conclusion is that deontological philosophy, rather than being a genuine exercise of reasoning, is a rationalization of deep-seated intuitions. In the next section, I am going to review some of the empirical evidence supporting the Central Tension Principle, while the fourth and fifth sections of this chapter are devoted to the presentation of the direct route and the indirect route of normative derivation respectively.

### 1.4. FMRI EVIDENCE FOR THE CENTRAL TENSION PRINCIPLE

Consider again the switch dilemma:

SWITCH DILEMMA

> "You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman.
>
> If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman.
>
> Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?"

Now consider this similar dilemma, which has come to be known as the *footbridge dilemma:*

> FOOTBRIDGE DILEMMA
>
> "A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large.
>
> The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved.
>
> Is it appropriate for you to push the stranger onto the tracks in order to save the five workmen?"

Most ordinary people, and even most moral philosophers, firmly believe that it is morally appropriate to hit the switch in the switch dilemma, and they have a similarly strong belief that it is morally inappropriate to push the stranger in the footbridge dilemma, even though, from a numerical perspective, the two dilemmas are identical. In other words, people usually give a consequentialist judgment in the trolley dilemma and a deontological judgment in the footbridge dilemma, even though in both cases it is five lives saved at the price of one. For Greene and colleagues (2001), this pattern of judgment constitutes a "a puzzle for moral philosophers" (p. 2105), which is known as the *trolley problem*.

There are two ways of tackling the trolley problem. One is trying to find a *normative* solution. This is usually done under the assumption that our beliefs in the two dilemmas are correct. Then, the philosopher searches for a principle which would make the apparently

contradictory pattern of response justified. However, none of the normative solutions to the trolley problem has been really successful because "for nearly every principle that has been proposed to explain our intuitions about trolley cases, some ingenious person has devised a variant of the classic trolley scenario for which that principle yields counterintuitive results" (Berker, 2009, p. 298). Greene and colleagues (2001) believe that the task of finding a normative solution to the trolley problem is hopeless. So, they have focused on a *descriptive* solution. The descriptive solution refers to a problem not for the philosopher but for the psychologist. Greene and colleagues (2001, p. 2106) articulate this problem as follows: "How is it that nearly everyone manages to conclude that it is acceptable to sacrifice one life for five in the trolley dilemma [the switch dilemma] but not in the footbridge dilemma, in spite of the fact that a satisfying justification for distinguishing between these two cases is remarkably difficult to find?" Greene and colleagues' goal has been to find a factor which would explain the psychological difference we manifest when facing these two dilemmas.

There are many possible differences between the switch dilemma and the footbridge dilemma which could explain why we judge differently in the two cases. Greene and colleagues have focused however on the *emotional reaction* elicited by either dilemma (Greene et al., 2001, p. 2106): "[…] the crucial difference between the trolley dilemma [the switch dilemma] and the footbridge dilemma lies in the latter's tendency to engage people's emotions in a way that the former does not." The hypothesis is that the differing emotional reaction explains the difference between consequentialist judgment in the switch dilemma and deontological judgement in the footbridge dilemma. The difference in response between the two dilemmas can thus be used as a testing ground for Central Tension Principle. The hypothesis that the deontological judgment in the footbridge dilemma is related to emotional processes while consequentialist judgment in the switch dilemma is related to cognitive processes yields two main empirical predictions:

a) That brain areas associated with emotional processes would be more active during contemplation of footbridge-like dilemmas as compared to during contemplation of switch-like dilemmas (Greene et al., 2001, p. 2106). I call this the *neural activity prediction*.
b) Someone who reaches a nonstandard verdict in the footbridge dilemma (finding it morally appropriate to push the stranger) would do so by having his cognitive processes override the emotional response that suggests that it is inappropriate to push the stranger. Thus, those who reach a nonstandard verdict in the footbridge dilemma should take longer to reach their verdict compared to those who reach the standard (deontological) verdict, but no similar reaction time effect should be observed in those who reach a nonstandard verdict in the switch dilemma (since

there is no emotional response to override) (Greene et al., 2001, p. 2106; Greene, 2008, p. 44). I call this the *reaction time effect prediction*.

Both predictions have held (Greene et al., 2001; Greene et al., 2004). As for the neural activity prediction, it was found that contemplation of moral dilemmas like the footbridge dilemma and other similar scenarios produced greater activity in various emotion-related areas (like the posterior cingulate cortex, the medial prefrontal cortex, the amygdala, and the superior temporal sulcus). Furthermore, it was also found that contemplation of moral dilemmas like the trolley dilemma and other similar scenarios produced greater activity in areas related to cognitive processes (like the dorsolateral prefrontal cortex and the inferior parietal lobe). As for the reaction time effect prediction, subjects who gave nonstandard responses to dilemmas like the footbridge dilemma took significantly longer to provide their verdict compared to subjects who gave standard responses. As predicted, no significant time response effect was registered in subjects responding to the trolley dilemma and similar scenarios.

Assume, like Greene and colleagues do, that the difference in our reaction to the switch dilemma and the footbridge dilemma is to be found in the different levels of emotional engagement they elicit. Which difference between the two cases explains the differing emotional reaction? Greene and colleagues (2001) had to make a choice about which factor of the footbridge dilemma caused a stronger emotional reaction. They hypothesized that the crucial difference was in the fact that in the footbridge dilemma subjects were asked to imagine performing an act of "up close and personal" harm (pushing the stranger on the tracks); while what they were asked to imagine doing in the switch dilemma, despite leading to the same consequences, did not involve "up close and personal" harm. The difference between "up close and personal" harm and impersonal harm can be summarized by the "ME HURT YOU" criteria, which Greene and colleagues (2004, p. 389) describe in the following way:

> "First, the violation must be likely to cause serious bodily harm. Second, this harm must befall a particular person or set of persons. Third, the harm must not result from the deflection of an existing threat onto a different party. One can think of these three criteria in terms of 'ME HURT YOU'. The 'HURT' criterion picks out the most primitive kinds of harmful violations (e.g., assault rather than insider trading) while the 'YOU' criterion ensures that the victim be vividly represented as an individual. Finally, the 'ME' condition captures a notion of 'agency', requiring that all the action spring in a direct way from the agent's will, that it be 'authored' rather than merely 'edited' by the agent."

Following the use made by Greene and his colleagues, I will call forms of harm which respect these conditions *personal harm*, and moral dilemmas which involve this kind of harm (like the footbridge dilemma) *personal moral dilemmas* (correspondingly, I will call all forms of harm which do not meet the ME HURT YOU criteria *impersonal harm* and the moral dilemmas involving this kind of harm *impersonal moral dilemmas*). What Greene and colleagues found was that in each footbridge-like dilemma (personal moral dilemmas) people tended to judge deontologically if compared to switch-like dilemmas (impersonal moral dilemmas). They also found that the areas of the subjects' brains related to the processing of emotions were preferentially recruited by personal moral dilemmas rather than impersonal ones.

The reaction time effect prediction was later rescinded by Greene (2009) after a reanalysis performed by McGuire and colleagues (2009). The problem is that Greene and colleagues included in their analysis problems which are not real moral dilemmas, as the following:

> HIRED RAPIST DILEMMA: "You have been dissatisfied with your marriage for several years. It is your distinct impression that your wife no longer appreciates you. You remember how she appreciated you years ago when you took care of her after she was mugged. You devise the following plan to regain your wife's affection. You will hire a man to break into your house while you are away. This man will tie up your wife and rape her. You, upon hearing the horrible news, will return swiftly to her side, to take care of her and comfort her, and she will once again appreciate you. Is it appropriate for you to hire a man to rape your wife so she will appreciate you as you comfort her?" (Greene et al., 2001, supplementary materials).

Greene and colleagues considered this a personal moral dilemma. People almost always judge it is not morally appropriate to hire the rapist. So, this is the standard verdict in this moral judgment. Following the fact that they considered standard responses in the footbridge dilemma to be deontological, Greene and colleagues interpreted the standard response to the hired rapist dilemma to be deontological. In general, they interpreted all standard verdicts to personal moral dilemmas to be deontological. The problem is that cases like the hired rapist are not real moral dilemmas. Differently from the footbridge case, there is no real consequentialist consideration that would favor hiring the rapist. The answer to the hired rapist dilemma is completely obvious and is reached fast by the subjects. If the responses to these non-dilemmas are considered deontological, deontological responses for all personal dilemmas are indeed faster than consequentialist responses. However, if the responses to cases like the hired rapist are not considered deontological and are excluded

from the dataset, there isn't any response time effect between responding deontologically and responding consequentialistically to personal dilemmas. What emerges from the reanalysis performed by McGuire and colleagues, is that it takes people more time to provide an answer to a personal dilemma, regardless of whether it is a deontological or consequentialist response, than to provide an answer to an impersonal dilemma (this can be explained by the fact that personal dilemmas are generally more challenging from both a deontological and a consequentialist perspective).

Consider the following dilemma (Greene et al. 2004, p. 390):

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

This is a difficult moral dilemma. People provide different answers and usually take a long time to find an answer. In other cases, answering a moral dilemma is far easier, with the outmost majority of people finding a consensus and providing their answer in a very short amount of time. One such cases is the *infanticide dilemma*, in which a teenage girl must decide whether to kill her unwanted newborn (most people say that such action is wrong). According to the dual-process theory of moral judgment, moral dilemmas like the crying baby dilemma are difficult because the strong emotional reaction they elicit, which stirs the judgment toward a deontological position (it is wrong to smother the baby), has to compete with the more cognitive evaluation of the situation, which strongly suggests a consequentialist approach (after all, smothering the baby might be right, given that everyone, including the baby, would die as a result of one's inaction). On the contrary, moral dilemmas like the infanticide dilemma are easy because the strong emotional response against killing one's baby dominates the weak cognitive case in favor of the action (Greene et al., 2004, p. 390-391). In other words, dilemmas like the crying baby dilemma involve an increased level of "response conflict", that is conflict between different kinds of representations for behavioral response (Greene, 2008, p. 45). Greene and colleagues (2004,

p. 393) thus predicted greater activation in the anterior cingulate cortex (ACC), an area which they hold to be linked with processing conflict between competing representations, in subject facing difficult moral dilemmas like the crying baby dilemma compared to subjects facing easy moral dilemmas like the infanticide dilemma. The prediction held (Greene et al, 2004). The crying baby dilemma is similar to the footbridge dilemma in that it asks the subject to imagine performing an act of direct violence and thus evokes a strong emotional reaction. However, the crying baby dilemma also invokes a strong cognitive response which, in some subjects, can override the emotional response. Greene and colleagues (2004, p. 394) thus predicted that classically cognitive brain areas would show an increased level of activity when subjects contemplate difficult moral dilemmas as compared to easy moral dilemmas, even though these difficult moral dilemmas evoke a strong emotional reaction just like the footbridge dilemma. This prediction too held (Greene et al., 2004).

The aim of brain imaging studies by Greene and colleagues was to show that different brain areas are recruited when subjects face different kinds of moral dilemmas. However, these studies lend support to the Central Tension Principle only if one interprets the activity the brain areas which are differentially recruited by the contemplation switch-like dilemmas vs. footbridge-like dilemmas in terms of a difference between areas devoted to cognitive processes vs. areas devoted to the production of emotional reactions. With this assumption made, Greene and colleagues' findings seem indeed to support the division of labor between emotional processes and the cognitive processes in moral decision-making which is implied in the Central Tension Principle. What was found is that purportedly emotional areas of the brain are preferentially recruited by deontological judgments, while purportedly cognitive areas of the brain are preferentially recruited by consequentialist judgments.

### 1.5. COGNITIVE LOAD EVIDENCE FOR THE CENTRAL TENSION PRINCIPLE

Another prediction which follows from the Central Tension Principle is that *increasing cognitive load* (like a concurrent digit-search task) or *reducing available response time* should selectively interfere with consequentialist judgments, as only this kind of judgment is supported by deliberate processes, which have more limited resources. When faced with dilemmas in which there is a contrast between deontology and consequentialism, subjects under cognitive load should show an increased response time or a decrease in the frequency of consequentialist responses specifically. The results confirmed that cognitive load caused an increase in response time for consequentialist judgments selectively (Greene et al., 2008). It should be noted however, that cognitive load didn't cause a reduction in the frequency of consequentialist judgments, as originally predicted by Greene and colleagues.

As for the reduction of available time, interesting data comes from experiments conducted by Trémolière and Bonnefon (2014). They hypothesized that the degree by which consequentialist judgments would rely on deliberate processes, and thus be sensitive to reductions in available time, would depend on the kill-save ratio present within the proposed moral problem. If participants are presented a problem where they have to choose if to sacrifice one person to save five (a 1:5 kill-save ratio), then reduction in available time would have a strong influence on the frequency of consequentialist judgments. If, on the other hand, participants are presented a problem where they have to choose if to sacrifice one person to save one million (a 1:1.000.000 kill-save ratio), then not only consequentialist judgments would be more frequent, but they would be far less sensitive to reductions in available response time. In other words, a high kill-save ratio would make consequentialist judgments more automatic, making them depend on intuitive processes rather than deliberate ones. Evidence collected by Trémolière and Bonnefon seem to have confirmed the hypothesis. It should be noted that a strict interpretation of the dual-process theory is incompatible with these results: killing to save lives remains a consequentialist response, regardless of the number of lives saved. So, under the strict interpretation, as consequentialist responses are related to deliberate processes, we should see a reduction in these responses when cognitive load is applied or response time is reduced, regardless of the kill-save ratio proposed by the problem.

Gürçay and Baron (2017) found no increase in consequentialist judgments after nudging participants via direct instruction to employ careful reasoning and also found a slight opposite effect to that predicted by the dual-process theory: under cognitive load, there was an increase in consequentialist judgments. If Greene and colleagues' dual-process theory is correct, we should see deontological judgments come first (as product of the faster emotional drives) and consequentialist judgments appearing later in subjects willing to employ the scarcer resources of deliberate reasoning. To test whether this sequential view of the relation between intuition and deliberation is right, Białek and De Neys (2016) investigated whether deontological responders were sensitive to response conflict. Conflict scenarios are those where consequentialism and deontology cue different answers (as the trolley dilemma), non-conflict scenarios are those where they agree that the same answer is correct (for example, asking whether it is right to sacrifice five people to save one). According to the sequential view, those who respond deontologically do not employ the resources of deliberate reasoning – rather than computing both responses and choosing between the two they only compute the intuitive deontological response. So, they should not be affected by the fact that their response is in conflict with consequentialist considerations. The authors found however that deontological responders tended to process

conflict dilemmas differently than no-conflict ones: the time they took for producing their response increased and their confidence in the response decreased, probably meaning that the deontological response was in conflict with other kinds of considerations. The increased doubt of deontological responders was still present even when reasoning's resources were suppressed by a concurrent cognitive load task (Białek and De Neys, 2017). This can be explained by assuming that, contrary to the claims of the traditional dual-process theory, fast processes can also compute consequentialist considerations.

Bago and De Neys (2019) used a two-response paradigm to test the validity of the sequential theory of the relation between deontological and consequentialist responses. In the paradigm, participants are presented with problems and have to provide a first quick response, whose intuitiveness is granted by the fact that the participants operate under a strict deadline and concurrent cognitive load. Then, participants are given more time to reflect and provide the final answer. The authors found that when participants gave the consequentialist response after reflection, this response was most of the times also present in the first intuitive answer. Indeed, it seems that type 1 can also generate consequentialist responses. Furthermore, Rosas and Aguilar-Pardo (2020) found that imposing extreme time pressure slightly increases utilitarian responses.

In conclusion, the results coming from the time pressure and cognitive load studies are inconclusive at best and may provide evidence against the Central Tension Principle. Greene and colleagues themselves (2008) could not confirm their hypothesis that cognitive load should lead to a reduction in the frequency of consequentialist judgments, and further studies seem to show how intuitive processes, which Greene considers to be tied to emotional drives, can compute consequentialist responses. This is consistent with the conclusions reached in the previously: while there might be some anatomical specialization of deontological vs. consequentialist judgments, it is far from clear if this division can be mapped onto a distinction between emotion and cognition.

# 2. DEBUNKING ARGUMENTS AGAINST DEONTOLOGICAL JUDGMENTS

> "One cannot give too many or too frequent warnings against this laxity, or even mean cast of mind, which seeks its principle [the principle of morality] among empirical motives and laws; for, human reason in its weariness gladly rests on this pillow and in a dream of sweet illusions (which allow it to embrace a cloud instead of Juno) it substitutes for morality a bastard patched up from limbs of quite diverse ancestry, which looks like whatever one wants to see in it but not like virtue for him who has once seen virtue in her true form."
>
> *Immanuel Kant, "Groundwork for a Metaphysics of Morals" (1785)*

Most philosophers follow Hume (1740/1978) in the idea that it is illegitimate to derive an "ought" from an "is", and Moore (1903/1959) in the belief that the natural is not subject to attributions of rightness or wrongness, and that performing such an attribution is a conceptually mistaken "naturalistic fallacy". As shown by the above quotation, Kant (1785/1959) was particularly adamant that morality and the empirical world are two completely independent realms, and that one cannot derive the moral law through scientific enquiry. The dual-process model of moral judgment proposed by American psychologist Joshua Greene aims at showing how conclusions about morality can be drawn from discoveries in moral psychology. Greene is convinced that Hume and Kant were wrong and that we can derive some interesting "ought" from contemporary moral psychology.

## 2.1. DIRECT ROUTE: MORALLY IRRELEVANT FACTORS

In fourth section of Chapter 1, we have looked at the switch dilemma and the footbridge dilemma as examples of impersonal moral dilemmas and personal ones respectively, and we have seen how Greene and colleagues explain why we tend to judge in a consequentialist manner when faced with an impersonal moral dilemma and in a deontological manner when faced with a personal moral dilemma because this latter type of dilemmas tend to evoke in us a stronger emotional reaction. Now consider another example of personal vs. impersonal moral dilemmas: the drowning child case employed as an example by philosopher Peter Singer (1972). Singer argued that we in the affluent world have a moral obligation to do much more to improve the lives of needy people. We have in fact the moral obligation to do our best to help people when they are in immediate trouble in front of us. For example, we have, and we *feel* to have, a moral obligation to help a child when it is drowning in a shallow pond in front of us. So, we have the same moral obligation

to spend less money on luxuries for us to help starving children in poor countries. Greene (2008, pp. 46-48) thinks that Singer's argument is puzzling: people who believe themselves to be moral, and which would insist that we indeed have an obligation to save a drowning child in front of us, nonetheless spend money on luxuries instead of donating them for impoverished people elsewhere. We seem not to *feel* the same obligation towards faraway people as we feel for people near us. But why should mere distance affect our judgment? Singer aims a shedding doubt on our ordinary moral judgment by noting how it is influenced by a factor which is *morally irrelevant* – namely, spatial distance.

Notice how the structure of Singer's problem is the same as the trolley problem. In both cases, we have two numerically identical scenarios (one life at risk in both of scenarios in Singer's problem, one life for five in both dilemmas of the trolley problem). In both cases, we feel a surge of emotion when confronted with one scenario rather than the other (we feel more emotionally drawn by the drowning child than the starving child in Singer's problem, and more emotionally drawn by the act of pushing the stranger than hitting the switch in the trolley problem). In both cases the surge of emotion leads us to a deontological position in one scenario rather than the other. We feel an emotion towards the drowning child, and this leads us to believe that we have the *duty* to help him. We feel no comparable emotion towards the faraway starving child and so do not believe in having a comparable deontological *duty* towards him. We feel an emotion towards the act of pushing a person down the tracks and this leads us to believe that we would violate his *right* by pushing him. We feel no comparable emotion towards the person on the other track in the switch dilemma and so we do not believe her to have a comparable deontological *right* we would be violating. This case shows perfectly the link between emotion, deontology, and a specific factor of the problem we are facing (it being personal harm or spatial distance). While philosophers have usually tried to provide a normative solution to the trolley problem or Singer's problem, Greene has focused on a descriptive explanation. The explanation is that the steer towards deontology is caused by a surge of emotions. These emotions react to the presence of non-moral features present in personal moral dilemmas (like personal harm or spatial distance) which are absent in impersonal moral dilemmas, thus allowing in those cases a more dispassionate cognitive analysis of the situation.

### 2.2. DIRECT ROUTE: THE EVOLUTIONARY ARGUMENT

But why would the presence of personal harm or spatial proximity evoke in us a stronger emotional reaction, making a moral dilemma personal? The reason is evolutionary. Greene (2008, p. 43) writes:

> "The rationale for distinguishing between personal and impersonal forms of harm is largely evolutionary. 'Up close and personal' violence has been around for a very long time, reaching far back into our primate lineage (Wrangham & Peterson, 1996). Given that personal violence is evolutionarily ancient, predating our recently evolved human capacities for complex abstract reasoning, it should come as no surprise if we have innate responses to personal violence that are powerful but rather primitive. That is, we might expect humans to have negative emotional responses to certain basic forms of interpersonal violence, where these responses evolved as a means of regulating the behavior of creatures who are capable of intentionally harming one another, but whose survival depends on cooperation and individual restraint (Sober & Wilson, 1998; Trivers, 1971). In contrast, when a harm is impersonal, it should fail to trigger this alarmlike emotional response, allowing people to respond in a more 'cognitive' way, perhaps employing a cost-benefit analysis."

Nature doesn't leave it to our faculty of cognition to deal with things which are essential for our biological fitness. For ecological reasons, the survival of the individual human being became dependent on the thriving of the group, creating a degree of mutual dependence far superior to that seen in any other social animal whose social group is constituted by non-kin. And yet, we have the ability of intentionally harm one another. So, it became essential for the thriving of human groups, and the survival of the individuals inside of them, to have some mechanism for controlling personal violence. As personal violence is evolutionarily ancient, we can expect this mechanism of control to be quite ancient and primitive too, predating our recently evolved human capacities for complex abstract reasoning. Just as it has endowed us with instincts related to food and sex, Nature has provided us with an efficient system against personal harm, in the form of strong emotions against the idea of engaging in it. The emotions against personal harm, as they are extremely beneficial from a biological perspective, are "alarmlike". These emotions took shape in the ancestral social environment and were sculpted by phylogenesis to ensure our survival in contexts where more time-consuming and resource-demanding forms of reasoning would have been detrimental. In contrast, impersonal harm fails to trigger alarmlike emotions, allowing people to respond in a more "cognitive" consequentialist way (Greene, 2008, p. 43). So, we judge deontologically in footbridge-like dilemmas because we have a strong emotional reaction to them, and we have that reaction because these dilemmas involve personal harm, and we tend to feel stronger emotional reactions to personal moral violations than impersonal ones because in the ancestral social environment it was biologically beneficial to

have a strong emotional reaction to personal harm. It was so beneficial indeed that we are endowed with a quite primitive emotional alarm against personal harm, an alarm which is automatically set off when we experience or just contemplate acts of personal violence; an alarm which steer us against personal harm in virtue of it being personal harm, *without consideration of whatever consequences it might have*.

The same holds true in the case of Singer's problem. Why does spatial proximity evoke in us a stronger emotional reaction? Because moral obligations for people in front of us are a kind of obligation our ancestors could have felt in the ancestral social environment. Feeling those obligations and responding to them might have brought important social and biological advantages. But under no circumstance were our ancestors in a position to help, or even to know, about people on the other side of the world. We feel to have a greater moral obligation towards a drowning child in front of us rather than towards a starving child in a foreign country because we have a stronger emotional reaction in the first case than in the second one. We have that reaction because the drowning child involves a spatially close moral obligation, and we tend to feel stronger emotional reactions to spatially close moral obligations rather than distant ones because in the ancestral social environment it was more beneficial to have strong emotional reactions to spatially close moral obligations (indeed it was possible to have emotional reactions only to spatially close moral obligations).

Not all emotions are primitive. Greene believes that even consequentialist reasoning needs some kind of emotional appraisal of the situation After all, one needs some kind of motivational drive in favor of, for example, saving five lives at the price of one. Greene (2008, p. 64-65) writes:

> "[…] I am not claiming that consequentialist judgment is emotionless. On the contrary, I am inclined to agree with Hume […] that all moral judgment must have some affective component, and suspect that the consequentialist weighing of harms and benefits is an emotional process. But, if I am right, two things distinguish this sort of process from those associated with deontology. First, this is, as I have said, a weighing process and not an "alarm" process. The sorts of emotions hypothesized to be involved here say, 'Such-and-such matters this much. Factor it in.' In contrast, the emotions hypothesized to drive deontological judgment are far less subtle. They are, as I have said, alarm signals that issue simple commands: 'Don't do it!" or "Must do it!' While such commands can be overridden, they are designed to dominate the decision rather than merely influence it."

Consequentialist judgments are driven by emotions allowing people to attribute different motivational values to various possible actions; but, once the value has been attributed to

each action, these emotions allow for a cognitive evaluation which weighs each action against each other based on its value. In this sense, emotions in consequentialist judgment function "like a currency" (Greene, 2008, p. 41), allowing to assign value to the objects of moral judgment before they are compared one with the other. This allows consequentialist considerations to be flexible, exactly like one would expect from processes based on neutral representations (that is representations dethatched from any immediate behavioral valence). Greene (2008, p. 64) writes that:

> "Consequentialism is, by its very nature, systematic and aggregative. It aims to take nearly everything into account, and grants that nearly everything is negotiable. All consequentialist decision making is a matter of balancing competing concerns, taking into account as much information as is practically feasible. Only in hypothetical examples in which 'all else is equal' does consequentialism give clear answers. For real-life consequentialism, everything is a complex guessing game, and all judgments are revisable in light of additional details. There is no moral clarity in consequentialist moral thought, with its approximations and simplifying assumptions. It is fundamentally actuarial."

On the contrary, emotion in deontological judgments functions as an alarm, simply prohibiting that certain actions be performed or mandating that other actions be performed, without the possibility of comparing the action with its possible consequences or with other possible actions.

Why are emotions involved in deontological judgments "alarmlike" while the emotions involved in consequentialist judgment are "like currency"? This is because characteristically deontological judgments are judgments about the inherent quality of a moral action, and this inherent quality (the inherent wrongness of pushing the person down the tracks in the footbridge dilemma, the inherent rightness of helping a child which is drowning in front of us) is something which depends on factors we only care about because they were extremely biologically relevant in the ancestral social environment, to the point where Nature endowed us of pre-reflexive emotional reactions to them. In other words, the emotions which lead to deontological judgments are those more "basic" emotional impulses which were positively selected prior to the development of sophisticated cognitive skills to allow for social regulation in contexts where complex reasoning would have been too demanding and slow. Greene (2008, p. 60) writes that:

> "[…] our most basic moral dispositions are evolutionary adaptations that arose in response to the demands and opportunities created by social life.

> The pertinent question here concerns the psychological implementation of these dispositions. Why should our adaptive moral behavior be driven by moral emotions as opposed to something else, such as moral reasoning? The answer, I believe, is that emotions are very reliable, quick, and efficient responses to recurring situations, whereas reasoning is unreliable, slow, and inefficient in such contexts."

These "recurring situations" are those our ancestors found in their social environment. It is sufficient for us to recognize to be in one of these recurring situations that our primitive emotional system is activated. This emotional system reacts to the simple presence of those factors (like personal harm or a drowning child) which constitute the evolutionary reason for why these emotions exist in the first place.

Why is it that stronger emotional reactions stir our judgment towards a deontological conclusion? In other words, why would alarmlike emotions be related to deontological judgments? Because deontological judgments are precisely the kind of "no-matter-what" judgments which an alarmlike emotional system would produce. I rationally value that saving five lives is better than preserving one life, but when I recognize that I have to perform an act of personal violence (like pushing the stranger down the tracks) my emotional system is set off by the simple recognition of the presence of personal violence, and it prevents me from conceiving the action as moral, *no matter the consequences of my inaction*. I rationally value the life of a starving child far away from me as equal to the life of a drowning child in front of me, but when I see the drowning child being close to me, my emotional system is activated by the simple recognition of the spatial closeness, and this gives me an impellent sense of moral obligation towards that child, while I feel no corresponding emotional pull (nor corresponding moral obligation) towards the starving child, *no matter the consequences of my inaction*. But the emotions explaining these differences in judgment are all dependent on the challenges we faced in the ancestral social environment. It really doesn't matter that the starving child is distant, it only matters to me for the way my species has come to attribute a moral value to spatial proximity, an attribution which depended on the needs humans had during our phylogenesis. The personal or impersonal nature of an act of violence, the closeness or distance of a moral obligation do not concern the dilemma we are facing, they are not resulting from an evaluation of its moral properties, rather they depend on our psychological constitution which in turns depends on purely contingent aspects of our phylogenesis. They are *morally irrelevant factors*. Greene (2008, pp. 69-70) writes that:

> "There are good reasons to think that our distinctively deontological moral intuitions (here, the ones that conflict with consequentialism) reflect the

> influence of morally irrelevant factors and are therefore unlikely to track the moral truth. Take, for example, the trolley and footbridge cases. I have argued that we draw an intuitive moral distinction between these two cases because the moral violation in the footbridge case is "up close and personal" while the moral violation in the trolley case is not. Moreover, I have argued that we respond more emotionally to moral violations that are "up close and personal" because those are the sorts of moral violations that existed in the environment in which we evolved. In other words, I have argued that we have a characteristically deontological intuition regarding the footbridge case because of a contingent, nonmoral feature of our evolutionary history. Moreover, I have argued that the same 'up close and personal' hypothesis makes sense of the puzzling intuitions surrounding Peter Singer's aid cases […], thus adding to its explanatory power."

If Greene's evolutionary hypothesis is correct, deontological judgments depend on alarmlike emotions which react to factors we only deem relevant due to morally contingent aspects of our phylogenesis. Deontological judgments do not track an independent, rationally discoverable moral truths, but ought their origin to human, all too human contingencies. This gives us an immediate reason not to attribute any normative significance to this kind of judgment.

### 2.3. DIRECT ROUTE: NORMATIVE CONSEQUENCES

What would follow from the fact that deontological judgments, as they are brought by alarmlike emotional reaction, respond to morally irrelevant factors? Consider the pervasiveness of rationalization in human psychology. Greene (2008, p. 35-36):

> "In one experiment, for example, people were asked to choose one of several pairs of pantyhose displayed in a row. When asked to explain their preferences, people gave sensible enough answers, referring to the relevant features of the items chosen—superior knit, sheerness, elasticity, etc. However, their choices had nothing to do with such features because the items on display were in fact identical. People simply had a preference for items on the right-hand side of the display (Nisbett & Wilson, 1977). What this experiment illustrates—and there are many, many such illustrations—is that people make choices for reasons unknown to them and they make up reasonable-sounding justifications for their choices, all the while remaining unaware of their actual motives and subsequent rationalizations."

We are good rationalizers of choices and judgments we actually made for reasons we are unconscious about. And often these unconscious reasons are emotional in nature. We feel the emotion, but we don't know what made us feel in that way, although we can come up with a perfectly sensical explanation referring to rational factors. For example, we might come up with a complex normative explanation to justify our deontological judgment in the footbridge dilemma, but (Greene argues) we have empirical evidence of the fact that the real reason why we judge deontologically in the footbridge dilemma is our alarmlike emotions. If Greene evolutionary hypothesis is correct, those emotions are reacting to a specific feature of the dilemma, namely personal harm. So, whatever normative explanation we come up with is just a rationalization of emotional impulses. What is worse, it is a rationalization of emotional impulses which are reacting to morally irrelevant factors.

Or consider Singer's problem. We might come up with many complex normative explanations to justify our preference for saving the drowning child over the starving child. These explanations will probably involve reference to deontological terms like right and duty. We rationalizers might come up with an ingenious way of explaining why we have a duty towards the drowning child, and no comparable duty towards the starving child. But (Greene argues) we have empirical evidence to conclude that this talk of rights and duties is just gibberish: the real reason we make a difference in the two cases is because we have an emotional reaction in one case, and we have no comparable emotional reaction in the other; and the reason for this difference in our emotional involvement is due to morally irrelevant features of our ancestral environment (in this case, that we could only react to spatially close moral obligations).

To make one last example: Immanuel Kant was opposed to masturbation. He devised a complex normative deontological explanation for why masturbation is morally unacceptable. But we can suppose that the real reason why he was opposed to masturbation was in the (culturally mediated) disgust he felt towards it. All the deontological explanation is just a rationalization of an emotion. This emotion is due to morally irrelevant features of Kant's cultural environment, rather than to features of our ancestral environment, but the argument still stands that deontological philosophy (the talk of right and duties) is just a rationalization of emotions responding to morally irrelevant factors. Greene (2008, p. 68, italics in the original) writes that:

> "[…] [according to deontologists] there is a complicated, highly abstract theory of rights that explains why it is okay to sacrifice one life for five in the trolley case but not in the footbridge case, and it *just so happens* that we have a strong negative emotional response to the latter case but not to the former. Likewise, […] [according to deontologists] there is a theory of duty

> that explains why we have an obligation to help Singer's drowning child but no comparable obligation to save starving children on the other side of the world, and it *just so happens* that we have strong emotional responses to the former individuals but not to the latter. […] The categorical imperative prohibits masturbation because it involves using oneself as a means […], and it *just so happens* that the categorical imperative's chief proponent finds masturbation really, really disgusting."

If we are able to show that deontological judgments are primarily supported by emotional processes (as the Central Tension Principle states), that these emotional processes respond to morally irrelevant factors (as the evolutionary argument goes), then from the pervasiveness of rationalization in human psychology we can deduce that deontological philosophy is a mere rationalization of morally irrelevant factors- If it were just shown that deontological philosophy is a rationalization of emotion, deontologists would still be able to argue that such emotions might be tracking independent, rationally discoverable moral truths. But if emotions only track morally irrelevant factors, that is, factors which depend on morally contingent aspects of our phylogenesis or our cultural environment, then one can argue that deontological philosophy does not track independent, rationally discoverable moral truths, and that thus all deontological philosophy is a misguided endeavor. This is arguably the strongest normative consequence of the dual-process theory proposed by Greene.

### 2.4. INDIRECT ROUTE: MODEL-FREE VS. MODEL-BASED LEARNING ALGORITHMS

Greene has also offered another argument, the *indirect route*, which does not employ the notion of morally irrelevant factors, and which does not target intuitions for their evolutionary origin, but more generally for the way they form. Greene does not believe that intuitions are categorically bad. Rather, he thinks that the usefulness of automatic-emotional processes and controlled-cognitive processes depends on the context. Our intuition is efficient but rather unflexible; while controlled processes are quite expensive in terms of cognitive resources and time but they allow for a more flexible kind of information processing (Greene, 2014). So, there will be instances where it is better to trust our automatic processes – that is, our intuitions. How can we determine when intuitions are trustworthy and when not? In a nutshell, assessing the limits of the faculty of intuition boils down to determining its origin.

Consider the dual-process model proposed by Fiery Cushman (2013). He thinks that the distinction between emotion and cognition is inadequate to capture the difference between deontological and consequentialist judgments. Consequentialist judgments in fact

require an emotional appraisal of the situation: for me to be moved to act, it is not sufficient that I recognize that saving five lives is *more* than saving one, I need to believe that saving five lives is *better* than saving one – a belief I can only acquire if there is some motivational force moving towards one option rather than the other. Furthermore, deontological judgments require a cognitive evaluation of the situation: what makes a given action inherently impermissible or inherently required will depend on some property of the action which I have to somewhat dispassionately take into consideration. Greene (2008) too believes that some degree of affection is present in consequentialist judgments, although he believes that the kind of emotion which is involved in consequentialist judgments is qualitatively different from the kind present in deontological judgments. Another option is to reject the Central Tension Principle and believe that the distinction between deontology and consequentialism can be captured in some other way.

Cushman (2013, p. 275) proposes that deontological judgments are preferentially associated with a process that assign value directly to actions, while consequentialist judgments are preferentially associated with a process that assigns value to their expected outcomes. Each of these two processes involves elements of cognition (identification of relevant properties of actions and outcomes) and affect (attribution of value to actions and outcomes). The proper distinction between deontological judgments and consequentialist judgments is not a matter of the quantity or quality of affect involved in them but rather a matter of the target of valuation: actions (in deontology) vs. outcomes (in consequentialism). As Cushman notes, the distinction between valuations focused on action and valuations focused on outcomes neatly complements the distinction between automatic processes and deliberate processes, while the relation between such distinction and the emotion/cognition distinction is more complicated. Cushman (2013, p. 276) writes:

> "previous research on action-based value representation associates this mode of decision making with habitual, automatic behaviors, while outcome-based value representation has been more often associated with effortful, controlled behaviors. […] In this respect, the action/outcome framework is a natural complement to the automatic/controlled framework. The relationship between the action/outcome framework and the reason/emotion [cognition/emotion] framework is more complicated. A virtue of the action/outcome framework is that it embraces the role of information processing and value representation in each of the two systems. In this respect, it denies the basic premise of the emotion/reason distinction. Yet, it also places its emphasis on the structural role of value representations in decision making, and in this sense it shares with the reason/emotion

> framework a fundamental concern with the relationship between knowledge, value, and computation."

Then, Cushman (2013, p. 276-282) proposes a parallel between action- and outcome-based value representations and two classes of learning algorithms: *model-free* reinforcement learning and *model-based* reinforcement learning. Reinforcement learning algorithms guide an agent's choices in an environment (deciding) and compute reward value obtained to specify the optimal policy, that is the choices leading to the maximization of reward in the long run (learning). They do so by representing value and by making computation which allow past experiences of reward to guide future choices. However, each class of algorithms (model-free vs. model-based) employs past experiences in a different way. This leads model-free and model-based algorithms to have structurally different targets of evaluation.

Model-based algorithms "learn by building a causal model of the world they occupy" (Cushman, 2013, p. 277). This model is an internal representation, which should be conceived as a decision tree, containing information about the probability of entering certain states given the choosing of a particular action, and of performing a certain action given the state one is in – with each state in the model being associated with an outcome (positive or negative reward). Model-based algorithms guide action by searching through the decision tree, representing links between actions, states and rewards. Using this procedure, the agent can simulate various action possibilities and select that maximizing reward. Model-based algorithms model everyday conscious reasoning: given several options, we simulate through imagination the likely outcome of each option, and we select that leading to the greatest reward. Having an explicit long-term representation of the links between actions, states and reward, model-based algorithms can take into consideration distal outcomes and make farsighted choices. Cushman (2013, p. 277) writes that the choices of model-based algorithms:

> "[…] can be farsighted, in the sense that they specify policies that require many actions to obtain a goal. […] can be flexible, in the sense that they can be recomputed at any time to reflect updates to the model. And […] can be goal-oriented, in the sense that the agent can specify a particular desired outcome and then compute the necessary sequence of steps to attain the specified goal."

The major drawback of model-based algorithms is their computational expensiveness: as the number of available states and actions grows, it becomes progressively harder to build an accurate model of all links between states, actions, and rewards.

Contrary to a model-based learner, a model-free learner does not carry a causal model of the world, thus it cannot take into consideration distal outcomes and is much more short-sighted. Rather than performing searches on a decision tree representing links between possible actions and future states, model-free algorithms only consider the value of immediately available actions. While model-free algorithms are, from a computational perspective, far cheaper than model-based ones, they struggle to attribute value representations producing optimal sequences of choices. Cheapness comes at the expense of accuracy and flexibility. However, there are two learning "tricks" which allow model-free algorithms to correctly update their action values and to even guide the agent towards long-term goals. One is *prediction-error learning*, which allows the agent to maintain a representation of the value of an action based on the average value it obtains performing that action. The agent does this by comparing a prediction of the value of the action with the actual reward it obtains by performing it, then updating the value representation of the action according to some internally represented function. In prediction-error learning is not a reward itself which promotes learning but the discrepancy between predicted and actual reward. Prediction-error learning allows an agent to represent the average value of a choice without remembering all the past outcomes of that choice. Prediction-error learning is useful not only for model-free algorithms, but also for model-based ones, as it is one of the tools by which this kind of algorithms construct a model of the world.

The second "trick" is *temporal difference reinforcement learning*, which allows model-free algorithms to compare value representations acquired at different stages in time. When employing temporal difference reinforcement learning, the agent considers actions leading to a reward as inherently rewarding by themselves. So, if pressing a lever is conducive to a reward (like food), the agent starts to treat pressing a lever as if it were itself a reward, so it can start to value actions which lead to the pressing of the lever rather than directly to food. This allows the agent to construct chains of actions, establishing a path which allows him to efficiently complete complex task (he turn right, so that he can then turn left, so that he can then reach and press the lever, so that he can then receive food). It does so without a representation of the reward which awaits it at the end of the task, and without knowledge of the fact that each step is only a means to obtain the final reward. Rather, the agent considers each intermediate action as endowed with an intrinsic value and, in each state, it just performs the action which leads immediately to the highest reward. Contrary to model-based agents, model-free ones are quite inflexible, as a local change to a specific value representation cannot be used, given the lack of a casual structure linking each value representation to the others, to adjust behaviors globally. While they lack the flexibility of reasoning, model-free algorithms can be used to model trial-and-error learning, as the agent

keeps track of the value representations that follow from each of his choices and is guided by these representations to perform complex tasks while being computationally light.

As reported by Cushman (2013, p. 279), one example of the difference between model-free agents and model-based ones is the devaluation procedure. A rat is trained to press a lever to obtain a food reward. When the rat has learned the association between the pressing of the lever and the reward, it is taken out of the apparatus and fed until it shows no more interest in food, then it is returned to the apparatus. It has been shown that sometimes the rat will continue to press the lever even though it is now rationally useless to do so (Dickinson et al., 1995). This is consistent with the implementation of a model-free algorithm: the rat has come to associate the act of pressing the lever as inherently rewarding, regardless of the original reward. Other times, the rat will pay no attention to the lever. This is consistent with the implementation of a model-based algorithm: the rat has a model of the relation between pressing the lever and obtaining a reward; given that the reward is no longer of interest, the rat will not press the lever. The devaluation procedure is an effective way to show that model-free algorithms can lead to different behavioral patterns than model-based ones, and that both kinds of strategies are available to the mind of a rat, going to constitute a dual-process apparatus. The devaluation procedure also shows that the action-based vs. outcome-based difference can be mapped onto the model-free vs. model-based distinction: model-free agents will come to value actions not for any future meaningful reward, but for the intrinsic value they have associated to them given the their previous trial-and-error experience; on the contrary, model-based agents will value actions for their usefulness to obtain a future reward, which they have represented in an explicit long-term model of their world.

While Cushman's overall argument is that the distinction between model-based and model-free algorithms is a better way to capture the difference between consequentialist judgments and deontological ones than Greene's theory (because it can account for the fact that both deontological judgments and consequentialist judgments employ cognition and affect), there is a useful way in which Greene's theory can exploit Cushman's dual-process rendition of moral judgment (as exemplified by Greene, 2017). Framing intuitive judgments in terms of model-free algorithms allows us to capture the limitation of these kinds of judgments. Greene (2017, p 5) writes:

> "First, intuitive decision-making is likely to fare poorly in a changing world. More specifically, intuitions do poorly when the causal relationships between actions (in context) and consequences differ between the world in which the intuitions were acquired and the world in which they are subsequently deployed. This is illustrated by the model-free rat whose

> string of habitual responses cannot easily adjust to a relocated cheese. Second, even in a stable world, affective learning does poorly when there is a mismatch between the values implicitly embodied in the training/learning process and the values of the agent. This point is illustrated by the rat who continually eats poisoned food because it has received no signal urging it to do otherwise."

If intuition is indeed modelled by model-free algorithms, then the capacity of intuitive judgments to be shaped by experience will be limited in specific ways, which will then constraint its usefulness to specific circumstances. In particular, in order to learn effectively, model-free algorithms need *good data*, that is a sufficiently representative number of cases to learn from, and a *good trainer*, that is a mechanism which provides feedback that is aligned with the values that the agent considers important (Greene, 2017, p. 4). The need for a good trainer and good data limit the viability of model-free algorithms to cases where the agent has had good trial-and-error experience of the subject matter, and where trial-and-error experience is enough to achieve knowledge of the subject matter.

### 2.5. INDIRECT ROUTE: FAMILIAR VS. UNFAMILIAR MORAL PROBLEMS

There are instances where trial-and-error experience of one's ancestors is represented in one's genes, providing innate knowledge of certain perils or opportunities. When facing a recurring problem of our evolution, intuitions might be a good cognitively inexpensive way to tackle the problem. The example Greene (2014, p. 714) makes is that of the fear of snakes. But the trustworthiness of intuition is not just limited to cases where we have an innate predisposition. Not all intuitions are evolutionarily innate, but they can also form out of cultural or personal trial-and-error experience. The example Greene makes is that of the fear of guns. Fire weapons didn't exist in the ancestral environment, so that fear cannot reflect a genetic predisposition, but one can have an intuitive fear of them out of cultural exposure received about the danger they represent. Still in other cases, an intuition derives from familiarity gained through personal trial-and-error experience. One might, for example, have an intuitive caution towards a hot stove because he happened to put his hand on it once. In these cases, not only is intuition trustworthy, but it shows a remarkable degree of interpersonal or intercultural variation, and so a degree of flexibility. The point remains, however, that while intuitions can be flexible, the way in which they are formed is rigid. While intuitions show cultural and personal variation, being somewhat influenced by individual learning mechanisms, the need for good data and a good trainer limits the viability of intuitions to cases where the agent has had a good trial-and-error experience of the subject matter, and where trial-and-error experience is enough to achieve knowledge of

the subject matter. But there are many cases where trial-and-error experience is not sufficient to behave in an effective manner. Greene's example (2014, p. 714) is driving a car. Our ancestors didn't have cars (there is no evolutionary familiarity), cultural familiarity with cars is not sufficient to learn to drive one, and personal familiarity is precisely what a new driver lacks. So, it would be a cognitive miracle if intuition was enough to learn how to drive a car. So, there are some kinds of problems which intuitions, insofar as it is correct to characterize them as model-free algorithms, cannot tackle. Greene (2014) defines familiar moral problems as those for which we have sufficient evolutionary/cultural/personal trial-and-error experience and unfamiliar moral problems as those for which we lack this experience. The question of the trustworthiness of moral intuition then becomes a question on which moral problems are familiar, and which are unfamiliar.

Greene (2017, p. 72) argues that the moral problems usually taken into consideration by philosophers are unfamiliar moral problems, in which intuition is set to fail. There are various methods by which we can tell apart familiar moral problems from unfamiliar ones. One is conflict. Moral problems can be divided into two categories: "Me vs. Us" problems, which involve a tension between personal selfishness and the interest of the community, and "Us vs. Them" problems, which involve a tension between groups of people with diverging moral views and intuitions. While "Me vs. Us" problems surely involve a level of conflict between personal inclinations and societal expectations, the intuitions we have in those cases are remarkably similar across both individuals and cultures. The reason is that we usually have evolutionary or cultural familiarity with this kind of problems. "Me vs. Us" problems are in fact general problems of coexistence within human groups, and we can expect human societies to have developed, over the course of natural or cultural selection, shared intuitions about them. The lack of conflict that "Me vs. Us" problems present at the level of intuitions is a good sign that these moral problems are familiar ones. On the other hand, "Us vs. Them" problems present a conflict between different human groups (them being cultural groups or subgroups within a culture divided by politics, religion and alike) which have different views on how social life should be organized and different intuitions about how to solve given moral problems. Examples of these problems include how to manage immigration, how to balance taxation with social expenditures, in which cases allow abortion and so on. The lack of agreement at the level of intuitions in "Us vs. Them" problems is a sign of the fact that these are unfamiliar moral problems, arising from circumstances in which we do not have evolutionary or cultural familiarity.

The abstract problems tackled by moral philosophers are problems where people are often divided into camps and where intuition does not suggest a clear universal response. But there is a further reason for which these kinds of problems should be considered

unfamiliar moral problems. The reason is that these problems, which are often used against consequentialism, are highly hypothetical and unusual, as exemplified by the fact that in them the usual relationships between actions and consequences are reversed. Take the footbridge dilemma as an example: here we are presented with a hypothetical scenario where an act of personal violence against an innocent person leads to positive consequences. Our faculty of intuition, which has evolved in an environment where such a case rarely occurs, strongly suggests that we ought not to perform such an act of violence. Intuition here is based on a model-free algorithm, which considers the act of violence per se, based on the repeated evolutionary and cultural trial-and-error experience suggesting that acts of violence against innocents lead to negative consequences. What intuition cannot, due to its computational constitution, consider is the weirdness of the footbridge dilemma. Such dilemma confronts us with a situation which is the opposite of what we are used to facing given our evolutionary and cultural experience. In this situation, violence against innocent people is the only way that leads to positive consequences. Greene concludes that, while we might trust intuition in other instances, we ought not to trust it in this particular case. The footbridge problem is structured exactly to be the kind of problem where our intuition misfires, because it puts us in front of a situation for which a model-free algorithm is set to fail, that is a situation where our deepest expectations about the relation between action and consequences are betrayed. If we accept this conclusion, then we have a reason not to trust our deep emotional-deontological intuition that the consequentialist response to the footbridge dilemma is wrong. If we trust the intuition in this case, we are mistaken because the footbridge dilemma is precisely the kind of problem where intuition is set to fail.

Greene (2017, p. 10) argues that many of the philosophical counterarguments to consequentialism are based on this kind of weird case, where the relation between actions and consequences is the opposite to what we could expect given our evolutionary and cultural experience. More generally, deontological/intuitive judgments are related to intuitions, and these intuitions are quite unflexible in novel and difficult situations, so we have a positive reason to trust consequentialist judgments in those situations. The implications of the indirect route argument are similar to the direct route: deontological judgements are intuitive, and people, philosophers especially, tend to rationalize intuition. Thus, we can expect deontological philosophy to be a rationalization of intuition. But these intuitions, as we have seen, are trustworthy only in a narrow number of cases. So, each application of deontological philosophy to unfamiliar cases will be a rejectable case of rationalization.

# 3. BEYOND THE EMOTION-COGNITION DIVIDE

In this chapter, two related theses are discussed. The first one is that one of the key distinctions made within Greene's dual-process model, that between emotion and cognition, cannot be so easily mapped onto another distinction, that between intuition and deliberation. The second claim is that the emotion-cognition divide cannot be properly applied to the brain, meaning that it is not so trivial to associate brain areas with either a cognitive or affective function. In the previous chapter, we saw how the fMRI evidence supporting the Central Tension Principle only lends support to it if one assumes the idea that emotion and cognition can be ascribed to specific brain areas. Here, I will try to show how this assumption is untenable by first focusing on a specific brain region (the amygdala) and then extending the results to the general relation between functions and areas.

## 3.1. THE STANDARD APPROACH TO MORAL DECISION-MAKING

The Central Tension Principle implies a *faculty psychology* view of psychological function. Faculty psychology is "the view that many fundamentally different kinds of psychological mechanisms must be postulated in order to explain the facts of mental life" (Fodor, 1983, p. 1) and is to be contrasted with the view that the mind is a generic and undifferentiated workspace. Each faculty is individuated by its characteristic operations, and providing an explanation in faculty psychology means giving an account of the operations of these distinct faculties, explaining how they jointly contribute to a given task. The idea that each faculty is *fundamentally different* from the others should be conceived as implying a kind of encapsulation of each faculty from the others. On the faculty psychology view, "cognitive faculties are independent, separate, and operate according to their own internal operations" and "having these properties is precisely what constitutes being a cognitive faculty at all" (Saunders, 2016, p. 254). In other words, faculties are internally cohesive psychological structures which can only interact at an output level, so to jointly contribute to a given task, while the inner functioning of each faculty is not influenced by any other faculty. Saunders (2016, 254) calls the idea that an explanation for moral judgment shall involve the ordered operations of two distinct psychological faculties, namely cognition and emotion, the *standard approach* to moral decision-making. It should be stressed that the problematic point here is not the adoption of faculty psychology (which is almost certainly right), but the adoption of the view that emotion and cognition are two distinct faculties.

While on one level, Greene and colleagues' hypotheses are innovative, functioning as a synthesis between the cognitivism of traditional cognitive science and the more recent sentimentalist reaction to it, on the other it has also been said that the Central Tension

Principle depicts "an extremely old picture of how the moral mind works" (Berker, 2009, p. 301), which Berker refers to has the "combat model" of the relation between emotion and cognition. Not only is the model of the mind emerging out of Greene's dual-process theory of moral judgment one of division of labor and competition between emotional reactions and cognitive processes, but it is one where priority is assigned to cognitive processes, as processes which are responsive to norms of rationality, while the emotions exert a disruptive influence. This latter idea has been referred to by De Caro, Marraffa & Vaccarezza (2021) as the "hierarchical", "pyramidal" or "Victorian" view of the mind. This view, which they describe in the context of a criticism of the dual-process theory of moral judgment, "postulates a gradual ascent from lower psychological levels (such as instinctive drives, tensions, and animal automatisms), through increasingly higher psychological levels, up to a vertex that is able to both impart order to this hierarchy of functions and coherently direct the 'noblest' functions that constitute rational self-consciousness" (De Caro, Marraffa & Vaccarezza, 2021, p. 35). The question of whether the dual-process theory of moral judgment adopts a combat model of the relation between cognition and emotion can be phrased in the question of whether the dual-process theory holds that moral judgment is the product of two fundamentally distinct psychological faculties, namely cognition and emotion. More succinctly, the question is whether the dual-process theory adopts the standard approach to moral decision-making. Saunders' own answer to this question is "yes". He writes (p. 255):

> "According to Greene, reason [cognition] and emotion are independent systems for coming to a moral judgment. Reason produces characteristically utilitarian moral judgments, and emotion produces characteristically deontological judgments (Greene, 2008; Greene et al., 2001). On this view, moral dilemmas occur (or are felt to occur) when these two independent systems produce conflicting moral judgments (e.g. reason produces a judgment that a certain action is permissible, while emotion produces a judgment that the same action is impermissible). When such conflicts occur, an overall judgment is arrived at through a conflict resolution system, though the precise details of this mechanism are not spelled out. Importantly, though, this model again explains moral judgment by the ordered operations of reason and emotion, only in this case the two faculties operate more competitively than cooperatively."

Indeed, Greene and colleagues assume that our apparently contradictory pattern of judgment in the trolley problem is due to a conflict between cognition and emotion. The Central Tension Principle fully expresses the idea of a division of labor between cognitive and affective processes: deontological judgments are arrived at through emotional reactions,

while consequentialist judgments are arrived at through cognitive considerations. When we stir onto a deontological judgment on a dilemma which we would otherwise judge in consequentialist fashion (as in the trolley problem), this is due to a prepotent emotional reaction which interferes with our cognitive processing of the situation. When we are undecided between emitting a deontological judgment or a consequentialist one (as in the crying baby dilemma), this is due to a direct conflict between our emotional inclinations and our cognitive reflections.

My claim here is not just that the dual-process theory of moral judgment proposed by Greene arrives at a conclusion which implies the standard approach, but rather that it has to assume the standard approach as a *necessary precondition*. More precisely, the predictions which follow from the Central Tension Principle only lend support to that principle if one assumes particular versions of the standard approach. The neural activity prediction, for example, has to assume that areas of the brain differentially recruited by deontological judgments vs. consequentialist judgments can be assigned to an emotional function or to a cognitive function. If this assumption were to be proven to be unjustified, from the fact that Greene and colleagues' experiments show that there is some level of specialization in the brain between deontological judgments vs. consequentialist judgments, one would still not be able to derive that deontological judgments are associated with emotion and that consequentialist judgments are associated with cognition (that is, one wouldn't be able to derive the Central Tension Principle). This is what I want to show in the following sections of this chapter.

### 3.2. EMOTIONAL DELIBERATION: THE CASE OF MOODS

Before looking at the possibility of assigning an affective vs. cognitive function to the areas involved in moral judgment, I want to focus on the first of the two claims I set out to defend in this chapter: that the emotion-cognition divide cannot be mapped onto the intuition-deliberation one. In general, this claim is quite easy to defend, as it is almost obvious that, in spheres different from morality, there are psychological processes which are fast and unconscious but nonetheless not immediately emotional or behaviorally valenced (think of syntactic parsing), while there are processes that take time and are consciously felt which have an emotional component (think of moods). As we saw in the previous chapter, in his 2008 paper Greene briefly considers moods as emotional states which are extended in time, before stating that he will focus on those affective states which, besides being behaviorally valanced, are also quick and intuitive. I will argue here that the possible objection coming from moods cannot be so easily dismissed. Moods are global emotional states, which have a significant time duration, and which influence the reasoning strategies of the subject experiencing them. The conclusion that can be drawn by examining

moods is that they breach the classification made by Greene and they constitute a form of emotional deliberation, which might sometime be morally relevant.

People usually take their affective states as a source of information. When an evaluative judgment is complex, people often employ the "how do I feel about it?" heuristic: rather than focusing on the features of the target to be judged, they focus on the affect that target evokes in them. Often though, we might mistake our preexisting mood state for the feeling we experience in reaction to the target, and thus our evaluative judgment is a direct consequence of the mood we are in. The "how do I feel about it?" heuristic is not, however, impermeable from reasoning: the impact of emotions and moods on an evaluative judgment is a function of how much emotions and moods are perceived to be informationally valuable. If the person attributes her current feeling to a source that is irrelevant to the evaluation of the target, the informational value of their feeling would be discredited, and the emotion or mood would have no effect on the evaluative judgment (Schwarz, 2002). For example, people report higher life satisfaction when good weather induces a good mood in them, and lower life satisfaction when bad weather induces a sad mood. However, when people are made to notice that the weather might be a plausible transient cause of their mood, they tend not to draw on their feelings in evaluating their life satisfaction. At the same time, individuals do not change their answer when they are asked about how they feel, meaning that making them notice that the weather might be a cause for their answer on life satisfaction does not alter their mood, but just how much that mood is taken into account in providing the answer to the question on life satisfaction (Schwarz & Clore, 1983). In short, subjects might use their feelings, whether lasting or transient, as the basis for evaluative judgments unless the diagnostic value of such feelings is put into question (Schwarz, 2002).

Moods have been found to have an effect on global reasoning and information processing strategies. This effect is because moods are, in the context-dependent manner discussed above, trusted as a source of information in evaluative judgment. The threat of a negative outcome or lack of positive outcomes often induces in us a bad feeling, and the prospect of positive outcomes often induces in us a good feeling. So, we have come to trust a happy mood as the signal of a benign situation, and a bad mood as the signal of a problematic one. People tend to think in a way that is appropriate to the situation they are facing as it is signaled to them by their mood. So, when they are in a bad mood, individuals will come to believe to be facing a problematic situation and will tend to apply the appropriate information processing strategy, adopting a more systematic, data-driven, and detail-focused reasoning strategy. On the other hand, when they are in a happy mood, individuals will come to believe to be facing a benign situation, and will come to apply the appropriate information processing strategy, adopting a top-down and heuristic reasoning

strategy, making larger use of preexisting general knowledge structures, with less attention to detail. This reasoning strategy is less effortful and allows us to spare cognitive resources (Schwarz, 2002).

Evidence of the effect of moods on information processing styles comes, for example, from stereotyping. When we have to form an impression of others, we might rely on detailed information about the target person or simplify the task by drawing on preexisting knowledge structures, such as stereotypes pertaining to the person's social category. Consistent with the idea that good mood increases reliance on preexisting knowledge structures, happy mood has been found to increase stereotyping (Bodenhausen, Kramer & Süsser, 1994). On the other hand, sad mood is consistently associated with a reduction in stereotyping and increases the use of information pertaining to the target individual (Bless, Schwarz & Kemmelmeier, 1996). The effects that mood has on reasoning strategies are dependent on the perceived informational value of the mood state: when there are reasons to believe that our feeling responds to factors which are irrelevant to the evaluation of the situation, the effect of mood on reasoning styles is strongly reduced (Sinclair, Mark, & Clore 1994).

The reciprocal effect of moods on reasoning (happy moods induce a top-down reasoning strategy, and sad moods induce a bottom-up reasoning strategy) and of reasoning on moods (the influence of moods on evaluative judgments and on reasoning strategies is diminished if the subject understands that the mood informational value is dubious) has implication for the idea that the emotion-cognition distinction can be mapped over the intuition-deliberation distinction. Moods are non-intuitive emotional states that have a prolonged influence over cognitive capacities like reasoning styles, and even on moral deliberation. This influence is such that one cannot ascribe a specific response to emotion and another to cognition, but rather the two are working in concert, as we shall see more in depth in Chapter 4. For the moment, it will suffice to say that if moods have a moral significance, it will show how the emotions involved in moral reasoning are not necessarily punctual in time, and thus intuitive.

### 3.3. REGIONS AND FUNCTIONS: BROCA'S AREA AND THE AMYGDALA

In 1861, the French physician Paul Broca reported his studies on patients with production aphasia and his discovery that their inability to articulate language was due to lesions to a brain region in the frontal lobe of the left hemisphere, an area corresponding to Brodmann Areas (BA) 44 and 45, which we currently call *Broca's area*. A century later, Broca's discoveries were inserted into a model of language processing according to which Broca's area is essential to language production, and generally to syntax, while Wernicke's area (BA 22) is essential to language comprehension, and generally to semantics (Geschwind, 1970).

It seemed like we had localized in the brain two central functions of language. However, later studies started to put pressure on this early picture of the relation of language to the brain. For one, it became apparent that language comprehension is not exclusively tied to Broca's area. Damage to just BA 44 and 45 does not result in a permanent and complete loss of the ability to articulate language, and patients with production aphasia have much broader lesions (Ardila, Bernal, and Rosselli, 2016). Considering recent research, it seems that language production is a much more distributed process involving many disparate and distant brain regions, many of which are also related to perceptomotor functions. On the other hand, Broca's area itself has many other functions beyond language, and is especially linked to action understanding (Binkofski et al., 2000; Nishitani et al., 2005). The earlier picture of a one-to-one correspondence between function and brain region has thus been challenged on a double level: for one, the function is realized by multiple areas; furthermore, each involved area is also related to a variety of other apparently unrelated functions.

This reinterpretation of the function of Broca's area is still compatible with an involvement of it in language understanding, but the overall function of the area is no longer seen as tied to language *specifically*. While Broca's area is surely also related to the processing of language's syntax, it seems that it does not react to complex and hierarchically organized linguistic sequences *in virtue of them being linguistic*. Rather, Broca's area seems to be related to the processing of complex, hierarchically organized sequences of events, whether they are sequences of actions (see the references above), of musical notes (Maess, 2001), or of words. Broca's area seems thus to be related to the processing of syntactic structures regardless of the specific format of the elements they are composed of. The interpretation of the Broca's area as a region specific to language comprehension or of it as a general syntax-processing area are clearly mutually incompatible. The reason is that Broca's area is recruited by language's complex and nested structures not in virtue of them being *linguistic* structures, but just a consequence of the fact that in language, just as in motion and elsewhere, there happen to be many complex hierarchically organized structures. It is important to understand the historical change here: Broca's area was considered to be the anatomical seat of a language-specific module, but now both the anatomical association and the general functional taxonomy have been put into question.

The reason I have introduced Broca's area is that I believe that a similar historical process of reinterpretation to that occurred in its case has also happened to the amygdala: we had a modular interpretation of its function as the site of a specific basic emotion (fear), then new findings have led us to give another interpretation, and the amygdala is now considered an essential component of the brain network that processes attention. The new interpretation is consistent with the previous findings that the amygdala is recruited by the

perception of threatening stimuli, but the new evidence changes the overall meaning of the function. Just like we previously believed that Broca's area implemented language comprehension, and now it appears that the area is related to the processing of complex hierarchically organized sequences of events, regardless of their format, it seems that the amygdala responds to the perception of threatening stimuli not *in virtue of them being fearful*, but just in consequence of them usually being stimuli that elicit attention. The interpretation of the amygdala function as a fear module is thus incompatible with the current interpretation of it as a general mechanism for attention and the identification of relevant stimuli.

The new interpretation of the function of the amygdala finds among its early proponents Sander, Grafman and Zalla (2003) who have directly criticized the interpretation of it as the seat of a domain-specific fear module. They write (p. 305): "consider the hypothesis that measures of amygdala activity reflect amygdala specificity in the processing of fear-related stimuli. In this case, a key criterion to verify this hypothesis would be to show that the differences obtained in amygdala activation for the processing of fear-related versus neutral stimuli are never found when comparing amygdala activation for the processing of non-fear-related versus neutral stimuli." In other words, if the amygdala is involved in many further cognitive processes beside fear, from the activation of the amygdala when the organism is confronted by fearful stimuli, one would derive only weak evidence that the amygdala is specifically involved in fear processing. As noted by Poldrack (2006), for the empirical evidence to really back up the claim that a function is implemented by a specific brain region, one should also see the correspondence going the other way around, with stimulation of the brain structure eliciting that function. The risk for the fear module interpretation is that the amygdala is involved in a variety of other functions, and that it is recruited by threatening stimuli only because there is some functional similarity between these stimuli and its actual function.

Öhman and Mineka (2001) have provided an interpretation of the amygdala as the anatomical site of implementation of the fear module. The fear module is a phylogenetically old device "for activating defensive behavior (e.g., immobility or fight-flight) and associated psychophysiological responses and emotional feelings to threatening stimuli" (Öhman and Mineka, 2001, p. 485). The reason why humans and many other creatures are endowed with a fear module is evolutionary. Behaviors serve specific biologically useful functions. Specific behaviors in an explanation of evolutionary psychology must be meaningful units from a functional-evolutionary perspective. According to Öhman and Mineka, fear as a psychological construct is anchored to defense mechanisms (like escape and avoidance) which promote survival in face of dangers like predators. The fear module is particularly

sensitive to stimuli that are biologically relevant because they are related to threats, and the reaction to the inputs induced by the fear module has proven, over phylogenesis, to be a better response than the alternatives in ensuring the survival of the organism and its offspring. For this reason, the fear module has a deep evolutionary origin and is shared by humans and many other creatures. The four characteristics of the fear module Öhman and Mineka (p. 485) focus upon are *selectivity* with regard to input, *automaticity*, *encapsulation* and *specialized neural circuitry*. Each of these features was shaped by evolutionary contingencies. These properties are not independent from one other, and just one of them would not be sufficient to identify fear as an evolutionary module.

Selectivity means that the fear module is particularly sensitive to stimuli that have been correlated with threatening encounters in the evolutionary past. Selectivity explains while it is much easier to condition an organism to feel fear towards given stimuli rather than others. As it is evolutionarily beneficial to activate defense mechanisms rapidly, the fear module operates on limited input, identifying critical stimuli with minimal neural processing, and is activated automatically, giving immediate priority to threatening stimuli. Like many phylogenetically ancient systems it is not under voluntary control and its operations are stimulus-driven. Informational encapsulation as a property of the fear module depends on the fact that it is phylogenetically more ancient than conscious thought. Öhman and Mineka (p. 485) write: "a module tends to run its course with few possibilities for other processes to interfere with or stop it. In particular, evolutionarily shaped modules will be resistant to conscious cognitive influences because their origin typically precedes recent evolutionary events such as the emergence of conscious thought and language." It is worth highlighting that the authors consider the influence going the other way around – the fear module influencing reasoning – mostly as a distorting influence. Finally, the fear module runs on a dedicated neural circuitry. The authors (p. 486) write: "At the neural level, an evolved module is likely to be controlled by a specific neural circuit that has been shaped by evolution because it mediates the functional relationship between ecological events and behavior. In the case of modules that are of ancient evolutionary origin, such brain circuits are likely to be located in subcortical or even brainstem areas". The biological evolution of the fear module is reflected in its neural implementation, as phylogenetically older functions are to be found in phylogenetically older areas. In particular, Öhman and Mineka hypothesize that the fear module is centered around the amygdala.

Before looking at the evidence in favor of a connection between the amygdala and fear, it is important to highlight the general structure of the reasoning employed by Öhman and Mineka, a strategy which can be extended to other basic emotions, and which reveals their general conception of emotions. Basic emotions are psychological devises whose functional

unity derives from a unitary evolutionary function that each emotion is tailored to. Given the phylogenetic age of emotions, these functions pertain to the organism's basic necessities. In simple organisms, these basic necessities are satisfied by the perceptuomotor apparatus, but in more complex creatures emotions provide a more flexible link between the perception of the stimulus and the motor response. Emotions serve as basic motivational drives, which are phylogenetically older and functionally simpler than higher cognitive abilities. This means that each individual basic emotion will share many features of Fodor's modules. As with any other module, the number of features each basic emotion shares will vary, but generally basic emotions will present functional specialization, a certain degree of specificity to input, automaticity, and informational encapsulation from central cognition. The phylogenetic oldness of basic emotions is reflected both in their early ontogenetic development and by their neural location in ancient regions at the back of the brain, which are responsible for functions shared between humans and many other animals. In this functional taxonomy, basic emotions can be thus opposed to central systems. Whether or not a modular organization can be envisioned for central cognition, it has properties which are different from basic emotions. Cognitive systems are more recent and are less widespread, being particularly developed in humans. They are acquired later in ontogenesis and are underpinned by a different set of neural mechanisms. The relation between emotion and cognition ought to be considered in terms of an input-output relationship, whereby the output of the emotion-producing limbic system is fed as input to the phylogenetically recent regions related to cognition. This idea is the anatomical version of what De Caro, Marraffa and Vaccarezza (2021) call the "hierarchical" or "pyramidal" view of the mind.

The relation of fear to the amygdala is supported by a set of studies showing that bilateral lesion to the human amygdala compromises the recognition of fear in the facial expression of others (Adolphs et al., 1995; Broks et al., 1998; Young et al., 1995). Brain imaging studies have also confirmed that the amygdala is recruited by the recognition of fearful faces (e.g., Phillips et al., 2001). However, activation of the amygdala seems not to be related just to fear, as the impairment to recognize emotion in faces covers all negative emotions, but not happy expressions (Adolphs et al. 1999; Fine & Blair, 2000). Damage to the amygdala in primates compromises the correct emotional response to stimuli, including a lack of fearful responses but also covering other kinds of emotional reactions (Amaral, 2002). To account for the growing evidence that the amygdala was related to the processing of all negative emotions, it has been proposed that the amygdala is the seat of module specific for all unpleasant stimuli (Paradiso et al., 1999). However, we shall now turn to the evidence that the function of the amygdala isn't just linked to negative emotions.

### 3.4. THE AMYGDALA AND ATTENTION

While studies reporting on the activation of the amygdala in case of negative emotional states can still be explained by a sufficiently enlarged version of the fear module interpretation, the evidence that the amygdala gets also recruited by *positive* stimuli and emotions is much harder to accommodate. In an fMRI study, Breiter and colleagues (1996) found a similar activation of the anterior amygdala when the subjects looked at both fearful and happy faces compared to a condition when they looked at neutral faces. Hamann and Mao (2002) found a similar result relating to the presentation of both negative and positive words compared to neutral words, while Garavan and colleagues (2001) found the same for negative and positive pictures. Zalla and colleagues (2000) found that the change in frequency of rewarding ("win") or aversive ("lose") words given as feedback to the participants doing a task modulates amygdala activation. Researchers have focused on the role of the amygdala in the identification of biologically valuable stimuli and found that the area is active during the perception of sexually appealing stimuli (Redouté et al., 2000) and of food when the person is hungry (LaBar et al., 2001). More generally, the amygdala seems to be involved in the processing of biologically relevant information, even when this information is not affectively valenced. For example, amygdala activation has been found during the processing of unknown (Dubois et al., 1999) or untrustworthy (Winston et al., 2002) faces, and the monitoring of gaze (Kawashima et al., 1999). More generally, the amygdala is involved in the evaluation of socially relevant stimuli (Adolphs, 2010).

The amygdala is an essential component for stimulus-reward learning. In particular, the amygdala is critical for associating stimuli with the value of rewards, and for attributing a positive value to former neutral stimuli in Pavlovian and instrumental conditioning. In classical conditioning, a conditioned stimulus (CS), which usually is a neutral stimulus (e.g., the ring of a bell) is paired with an unconditioned stimulus (US), which is biologically potent (e.g., food). The organism presents a natural unconditioned response (UR) to the US (e.g., salivation). After the conditioning, the same response becomes also automatic in presence of CS, and thus it had become a conditioned response (CR). In modern cognitive science, it is believed that the CR passes through a mental representation of the US when the CS is presented. In fact, the CR is often sensitive to the current value of the US. Responding to the CS is sensitive to post-training alterations in the value of the US. Hatfield and colleagues (1996) had conditioned rats to tone-food pairings, then devaluated the food through a toxin, in the absence of the tone. When the tone was presented again, rats showed a spontaneous reduction in their CR. However, rats with damage to the basolateral amygdala failed to show sensitivity to the post-conditioning devaluation of the conditioned stimulus. They too had no more interest in the food, but presented the CR in the presence of the CS. This shows

how the basolateral amygdala is essential to the representation of the US, which allows for flexibility in associative learning.

A similar result was obtained with monkeys (Malkova, Gaffan, and Murray, 1997). Rhesus monkeys learned to discriminate pairs of objects and received food as a reward. The animals were then satiated with one of the two food types they received as a reward. Monkeys were then required to choose between two objects, each of which had been paired with one of the food items, and chose the object related to the food they were not satiated with. Monkeys with lesions to the amygdala instead chose randomly between either type of object, even that which was associated with the food they weren't currently interested in. In another experiment in the same paper by Hatfield and colleagues (1996), rats were first conditioned through light-food pairings, then they experienced tone-light pairings. Rats acquired a second-order CR to the tone, but rats with damage to the basolateral amygdala did not: the light failed to acquire a reinforcing value based on its first-order pairing with food. These experiments show how the amygdala is necessary for the representation of the value of rewards.

It should be noted that not all value attribution requires the amygdala. Animals with damage to the amygdala or even with a complete removal of it have no problems with first-order classical conditioning and present an almost normal set of preferences, showing that they correctly attribute value to positive and negative stimuli (Baxter and Murray, 2002). What the amygdala seems central to is the *representation* of value. The rats that had received conditioning with a certain food, that was then devalued, presented a reduced CR when confronted by the CS that was associated with that food, most probably because they stored an internal representation of the value of the US. Such representation allows to rapidly update the value of the US, and consequently that of the CS, in face of events that change its benefit for the organism. Evidence for the role of the amygdala in a trans-modal and global representation of value comes from a study by Belova, Paton, and Salzman (2008). They performed a trace-conditioning study on rhesus monkeys by pairing three visual CSs to three USs who varied in magnitude, sensory modality, and valence. Then they presented the stimuli to the monkeys and recorded the activations of neurons in the basolateral and central amygdala during the trace period, when the animal came to expect a reward or a punishment. They noted that individual neurons in the recorded areas tracked the positive or negative value of the current state. Some neurons increased their firing more strongly after a positive CS and after the administration of the reward than after a negative CS and the administration of punishment, while other neurons did the opposite. Thus, neurons in the amygdala tracked moment-to-moment changes in the value of the stimuli. Interestingly, the experimenters also found that the amygdala provides a *graded* representation of value:

in a positive-value-coding neuron responses were high if a large reward was expected, low if a punishment was expected, and intermediate if a small reward was expected. In a negative-value-coding neuron responses were high on trials where a punishment was expected, low if a large reward was expected, and intermediate on small reward trials. The responses to the CSs thus reflected the integration of information about multiple reinforcers in different sensory modalities. This suggests that the amygdala is crucial for associating CSs and USs when these differ in sensory modality, valence, and magnitude.

The amygdala is not directly involved in the *motivational* aspects of value attribution, but in its *attentional* aspect. The minds of humans and animals must continually process vast amounts of incoming sensory information and determine which stimuli are worth attending at any given time. Attention is a complex and hard-to-define concept which refers to the capacity of the mind to highlight some stimuli while discarding others. Attention is ubiquitous in cognitive processing and is probably realized by a vast array of different functions. One aspect of attention is the ability of an animal to interrupt current behavior and orient himself toward a novel or unexpected stimulus. Early studies revealed that stimulation of the amygdala cause the arrest of current activities and the production of "alerting" and "searching" behaviors (raising the head and looking in an inquisitive manner), which can be interpreted as orienting movements (Kaada, 1951). Stimulation of the central nucleus of the amygdala also caused desynchronization of the cortical EEG (Kapp, Supple, and Whalen, 1994), which has been long considered a sign of cortical arousal (Moruzzi and Magoun, 1949). Attention is affected by the novelty and unexpectedness of stimuli. Novelty also modulates amygdala responses. Blackford and colleagues (2010) reasoned that unexpectedness can come in two forms: either objects are ordinary, but novel in the current context; or they are unusual. They selected pictures to fall within three categories: familiar common pictures, novel common pictures, and novel uncommon ones. The pictures were selected so to exclude emotional or social relevance, and participants were subject to the fMRI while looking at the pictures. What was found is that novel common pictures engaged both the hippocampus and the amygdala more than familiar pictures, while only the amygdala registered a higher level of activation when confronted by novel uncommon pictures than novel common ones, meaning that the amygdala is particularly tuned for the identification of both types of unexpectedness.

Temporal unpredictability also influences amygdala responses. Herry and colleagues (2007) exposed mice and humans to a simple repeating tone. In one condition the tone was part of a predictable sequence, while in another condition it was part of an unpredictable sequence. The fMRI revealed greater activation of the amygdala during the unpredictable vs. predictable condition in both mice and humans. Expectation and surprise modulate

amygdala responses. Belova and colleagues (2007) measured amygdala responses to reward and aversive stimuli in monkeys in two conditions: in one case the stimulus was predicted by a CS, while in the other case the stimulus was unpredictable. It was found that neurons in the amygdala responded more strongly when the reward or punishment violated their expectations. Some neurons exhibited differential responses to reward only and responded more strongly when the reward occurred unexpectedly. Other neurons exhibited differential responses to punishment only and responded more strongly when it occurred unexpectedly. Other neurons still responded more strongly to unexpected stimuli than expected ones but did not differentiate between valence. The amygdala seems thus involved in the modulation of reinforcement value given the expectations of the organism. Surprise by itself, regardless of the valence of the surprising stimulus, is enough to provoke a stronger activation of the area.

Selective attention shapes perception, by making the organism focus on those visual items which are considered most relevant. The process of stimulus selection is often influenced by the emotional value attributed to stimuli. Lim, Padmala, and Pessoa (2009) investigated how affective significance influences visual perception during an attentional blink task. In this kind of tasks, the subject is confronted with a rapid stream of visual stimuli and has to identify and report on a set of specific targets. If a target follows another by a brief delay, the subject is more likely to miss it. In the first phase of the experiment, images of houses or buildings were paired with a mild shock, so become affectively significant for the subjects. Subjects showed better performance in detecting affectively significant images (those of houses or buildings) than non-valenced stimuli in the attentional blink task. The amygdala and brain cortex responses were also stronger in the trials employing the emotionally relevant images. Even though the increase in the amygdala activation was predictive of better behavioral performance, the relationship was no longer statistically significant once the influence of the visual cortex was taken into account, meaning that whatever effect the amygdala had on behavior was mediated by the visual cortex. The authors were also interested in how the amygdala helps shape perception. They reasoned that if this is the case the strength of the predictive effect between activation of the visual cortex and behavioral performance should depend on signals coming from the amygdala. To test this, they performed a mediation analysis involving the amygdala, the parahippocampal gyrus, and the subject's performance on individual trials. They found that the strength of the effect of visual cortex to behavior (measured through the slope of the logistic fit) was correlated with the magnitude of evoked responses in the amygdala: when amygdala responses were weak, the relationship between the visual cortex and behavior was also weak; when they were strong, the relationship was also strong. This finding points

to a crucial role of the amygdala in orienting perception and determining which stimuli are worth attending at any given moment.

The evidence reported so far hints of the role of the amygdala in decision-making. Decision-making requires the evaluation of costs associated with each possible action in relation to their potential rewards. One paradigm used to study animal decision-making involves contrasting a possible action with immediate but small reward to one with delayed but larger reward. Animals with lesions to the basolateral amygdala show increased preference for the immediate reward. This pattern of choice is probably a consequence of the fact that they cannot form a representation of the value of the reward, that would allow them to consider such value in the absence of the reward (Winstanley et al., 2004). Rats with lesions to the basolateral amygdala also show a reduced tolerance for risk, preferring a small but certain reward to a larger but uncertain reward (Ghods-Sharifi, Onge, and Floresco, 2009). Generally, damage to the basolateral amygdala has been linked to more impulsive and risk-adverse behavior. In humans, these deficits manifest as generally compromised social behavior: patients tend to act more impulsively and to make worse choices than healthy subjects. For example, van Honk and colleagues (2013) found that amygdala lesions subjects were more trustworthy of unfamiliar persons compared to healthy subjects.

We can conclude here that the role of the amygdala is not just to produce fear, intended as a modular function. Rather the area seems to be a crucial component of the system directing attention and deciding which stimuli is worth prioritizing at any given time. The conclusion is thus similar to that reached in the case of Broca's area. We previously believed that the area implemented language's production, and syntax in particular, but now it appears that it is related to the processing of complex hierarchically organized sequences, regardless of the format of their components. Similarly, the interpretation of the amygdala's function as a fear module or as a negative emotions module is incompatible with the current interpretation of it as a general relevance detector: it seems that the amygdala is recruited by the perception of threatening stimuli not in virtue of them being fearful, but just in consequence of them usually being stimuli that elicit attention. It should also be stressed that the new interpretation of the amygdala assigns it a function which cannot be framed as purely emotional or purely cognitive. While the processes of selecting which stimuli to attend and to assign value to rewards have an essential emotional component, they also require a complex evaluation of the many factors influencing decision-making. As we have seen in the review of empirical evidence, the amygdala produces a representation of value which integrates various sensory modalities and which can be stored for a prolonged period, allowing for a flexible relation between stimuli and the decision-maker. Indeed, damage to the amygdala reveals that the area is involved in the making of choices which

can be properly defined as cognitive, and which can be contrasted with more impulsive and emotional behavior. Choosing a larger but delayed prize over a smaller immediate prize can be thought of as the cognitive choice, and both mice and human patients with damage to the amygdala have troubles with this kind of choices. For short, the function of the amygdala is misunderstood if it is analyzed in terms of a functional taxonomy which divides cognition and emotion and applies the distinction to the brain.

### 3.5. NEURAL REUSE THEORIES

A theory within moral psychology posits that emotional reactions are related to deontological judgments. Proponents of the theory claim that there is evidence that emotional brain areas are preferentially recruited by deontological judgments. My reply to this idea concerns the fact that, as we have seen for the amygdala, is far from clear if "emotional" areas of the brain can be identified. More generally, I would argue that it would be surprising to find that certain brain areas are exclusively related to one or more basic emotions, as the brain is not hierarchically organized with phylogenetically older areas subserving phylogenetically older functions. This last section can be seen as a generalization of the conclusions concerning the amygdala we have reached in the previous one.

A *complex system* is a large network of relatively simple components with no central control, exhibiting emergent complex behavior, meaning that the global behavior of the system arises from the collective action of single components, with a non-trivial mapping from individual actions to the collective behavior. Examples of complex systems include the immune system, ant colonies, the Internet, economic markets, and the human brain. *Network thinking* is a particular way to analyze complex systems, which is directly opposed to the decompose-and-localize method typical of modular approaches. According to the latter, the overall function of the system is broken into subfunctions, each of which is then associated with one of the physical subparts of the system. On the contrary, network thinking suggests that we should look for higher-order patterns in the behavior of the complex system in its entirety and employ those to explain the functioning of the system (Mitchell, 2006). An application of network thinking to the brain is represented by *neural reuse theories* (Anderson, 2010), which are directly opposed to modular approaches to the mind.

The basic idea of neural reuse theories is that there are evolutionary considerations to conclude that existing physical components are reused for new functions rather than developing new circuits de novo (Anderson, 2010, p. 246). This reuse can be conceived in terms of an exaptation of established neural circuits for new purposes. However, neural reuse only partially fits the exaptation account, as most often reuse does not entail a loss of the original function. At least three predictions follow from this idea. The first prediction is that a typical brain region will support numerous cognitive functions in diverse task

categories. Second, we should expect a correlation between the phylogenetic age of a brain area and the frequency with which it is redeployed in various cognitive functions. Phylogenetically older areas, that have been available for reuse for longer, are generally more likely to have been integrated into later-developing functions, and we should see that they are recruited for a greater variety of functions compared to more recent areas. Third, we predict a correlation between the phylogenetic age of a function and the degree of localization of its neural components. The more recent functions, like language, are more likely to be distributed over a greater amount of brain regions and those regions are probably more widely scattered than the regions supporting phylogenetically ancient functions (*ibid.*). These predictions are the opposite of what theories adhering to anatomical modularity should predict. If these predictions were proven right, they would have significant implications for our conception of the relation between cognitive function and anatomical structure. Rather than a functional architecture whereby individual regions are dedicated to large-scale cognitive domains like vision and language, they imply a picture of the mind where neural circuits are used for various purposes in different cognitive tasks over a broad spectrum of domains (*ibid.*).

Anderson (2007a, 2007b, 2008) has found evidence supporting the three predictions. As for the first prediction, it was found that the typical brain region is involved in functions related to *nine* task domain, which often included action, vision, audition, attention, emotion, language, mathematics, memory, and reasoning. To test the second prediction, Anderson made the simplifying assumption that phylogenetically older brain regions lie at the back of the brain. He found a negative correlation between the position of the brain region along the Y-axis on Tailarach (a tridimensional human brain atlas) and the number of tasks that region was involved in, showing how older regions were involved in a greater amount of tasks (the correlation is negative because the origin in Tailarach space is located at the center of the brain and posterior regions are increasingly negative). The third prediction is supported by the fact that tasks related to language, a phylogenetically recent function, seem to activate more regions than tasks related to older functions like visual perception and attention, and those regions are more widely scattered in the brain.

One of the main sources of evidence for neural reuse is the reuse of motor and perceptual areas for tasks related to higher cognitive functions. As noted by Zerilli (2019), these findings make it hard to argue for a distinction between central and peripheral systems, which was one of the main reasons motivating Fodor's modular approach. Evidence suggests that peripheral systems, underpinning perception and motor control, are recruited in various cross-domain tasks, including those related to central reasoning processes. Early evidence in this direction came from studies on lexical retrieval showing

how retrieval of words for selective categories of entities could be selectively damaged – with some patients being able to retrieve nouns but not verbs (Damasio and Tranel, 1993) or nouns of animal species but not those of tools (Damasio et al., 1996). The studies reported how the generation of color words recruits areas related to color perception (like the ventral temporal lobe), while the generation of action words recruits areas related to the perception of motion (like the middle temporal gyrus) (Martin et al., 1995). Retrieving names of animals recruited the left media occipital lobe – a region involved with visual processing -, while naming tools is associated with motor control areas (Martin et al., 1996). Similarly, the processing of food-related concepts recruits gustatory areas (Simmons, Martin, and Barsalou, 2005), and when people are confronted with concepts of things that smell, olfactory areas become active (Gonzáles et al., 2006). Today, the evidence for reuse of areas of the brain related to perceptuomotor tasks in tasks related to language, reasoning, and the processing of concepts is extensive (for some reviews: Barsalou, 2008; Kiefer & Pulvermüller, 2012; Meteyard et al., 2012; Dove, 2016). Although the precise implications of this evidence for the status of concepts as abstract or concrete entities are disputed and depend on the criteria which are employed to evaluate the empirical evidence (Raia, 2023), the evidence seems conclusive that there is some degree of reuse of phylogenetically old perceptual and motor areas of the brain for later-developed functions.

The consequence of adopting a neural reuse interpretation of the relation between function and brain area is that, contrary to traditional explanations in evolutionary psychology, we should not expect phylogenetically older areas to be linked to more "primitive" functions. More generally, the neural reuse interpretation speaks against the idea that cognition and emotion are two distinguishable sets of functions, with identifiable and distinct neural underpinnings, that interact at an input-output level. As we have seen in the previous sections, the amygdala was considered the seat of the fear module, but its function covers a much broader set of emotional and cognitive aspects. Simply put, neural architecture does not recapitulate phylogenesis. Not only the function of single areas but also the overall architecture of the human brain supports this conclusion. For example, it can be noted how the distance, in terms of connection among cortical areas, between areas traditionally considered to be affective and the sensory periphery is equivalent to the distance between prefrontal areas, traditionally associated with higher reasoning, and the periphery, meaning that both areas receive highly processed and integrated sensory information and suggesting that both are involved in the production of high-level responses to the environmental stimuli, with affective areas like the amygdala and the hypothalamus playing a crucial role in the integration of information (Pessoa, 2008). Further evidence comes from the fact that recent measurements have found that, contrary to the traditional

assumption, human frontal lobes are not relatively large compared to the rest of the brain (Barton and Venditti, 2013). Rather than searching for human uniqueness in cognitive capacities that have their seat in the frontal lobes, we should look at the interaction between regions and networks. Indeed, it has been proposed that human uniqueness and the emergence of language are related to the globular rather than elongated shape of our skulls, that allowed for closer and faster interaction between regions (Boeckx and Benítez-Burraco, 2014).

# 4. THE EDUCATION OF INTUITIONS

> "Man has a far greater variety of *impulses* than any lower animal; and any one of these impulses, taken in itself, is as 'blind' as the lowest instinct can be."
>
> - *William James, "Principles of Psychology" (1890)*

## 4.1. THE PHYLOGENETIC STANDARD APPROACH AND ADDITIVE THEORIES OF RATIONALITY

Consider the evolutionary hypothesis of Greene's direct route argument. Deontological judgments derive from emotions. Emotions are a quite primitive alarmlike system which reacts to factors that were important in our phylogenesis. So, emotions do not react to morally relevant factors. Such factors can only be appreciated with dispassionate cognitive analysis. A good part of our emotional reactions are "basic" emotional reactions, which function as a quite primitive alarm, just like our fear of snakes is set off by our simple perception of snakes and commands us quite primitive behavioral reactions to them. These emotional reactions can be intervened upon by cognitive processes (after all, there are people who keep snakes as pets), but this is an *exogenous* intervention, which only takes once the emotional system has provided its output. Insofar as their inner processes are concerned, these "basic" emotions, just like basic instincts for food and sex, cannot be *educated* by conscious processes, after all they are made to react to ancestral problems. Just as fear of snakes is an instinctive and innate reaction, which is designed to keep us alive in face of an ancestral biological threat, so the basic emotions which give rise to deontological judgments are an instinctive and innate reaction, which is designed to keep us alive in face of ancestral social threats. Just as fear of snakes is not something which can be properly attributed to any specific individual, but rather is a fixed species-specific reaction, so too are the basic emotions behind deontological judgments. While higher processes can diverge from individual to individual, but "basic" emotional intuitions, just like other efficient systems, are universal in the species.

If my reconstruction of Greene's evolutionary argument is correct, this should be the view of basic emotions which emerge out of the argument. The emotions which give rise to deontological judgments might be a bit more refined than fear of snakes, but they belong to the same general psychological category. I will call the idea that emotions (excluding "currency-like" emotions involved in consequentialist judgment) are primitive alarms made to react to ancestral threats, and thus that they are species-specific and universal processes, encapsulated from central cognition, the *phylogenetic standard approach*. I view this

conception of emotion as a form of standard approach, because it posits a difference between affective processes and cognitive processes. This difference is at the level of their phylogenetic origin: while emotions are phylogenetically ancient (and, for the most part, shared with other creatures), cognitive processes are more recent (and, for the most part, typically human). The evolutionary argument by Greene rests on this assumption. Deontological judgments are not a good moral guide because they are responsive to morally irrelevant factors, and they are so because the emotions originating them were shaped by phylogenesis to always and automatically react to certain morally contingent factors. If one does not assume that emotions are somewhat fixed and automatically reacting to given morally contingent factors, there would be no way of making sense of Greene's argument against deontology. So, we arrive at the conclusion that the normative implications of the dual-process theory of moral judgment proposed by Greene necessarily require the adoption of the phylogenetic standard approach.

The phylogenetic standard approach can be framed in terms of *additive* theories of rationality (Boyle, 2016), according to which being a rational animal can be thought of as constituted by two components: first, we are endowed with a primordial system that is shared with non-rational creatures; second, we have a reasoning system monitoring and regulating the activity of the primordial system. The primordial nature of the first system is both phylogenetic – activities like perception, volition, and basic emotion developed in the distant past and are thus common to many species -, and anatomical – we can suppose that phylogenetically more ancient functions are implemented by more ancient parts of the brain. Both a phylogenetic and anatomical additive view of rationality are implicit in and necessary to Greene's dual-process theory. Anatomically, as we saw in the previous chapter, there are regions of the brain that implement emotional states which can be differentiated from more recent regions which implement cognitive processes (on this crucial distinction rests the possibility of verifying the Central Tension Principle). Phylogenetically, the emotions which correlate to deontology are basic emotions which react to problems in the ancestral social environment. These emotions are primitive and are of the kind other creatures can experience. Being human means experiencing these basic emotional states on the one side *and then* having a deliberative system which takes the product of the basic emotions as inputs and oversees them. The deliberate system can suppress the emotional response or rationalize it, or it can produce an alternative cognitive response and so on. The point is that the two dimensions, while interacting, are not integrated, and it is possible to distinguish between the two functionally, anatomically, and phylogenetically. Additive theories are linked with a broadly Cartesian worldview: emotions (just like perception and basic volution) come from what is shared between humans and other animals, while

cognition is properly and exclusively human and is a further capacity added to the emotions; echoing the Cartesian idea that we share with animals the possession of a body, but we alone have a soul, which interacts with the body moderating the passions which originate from it.

A different, *transformative* view of rationality can be traced back to Aristotle. For Aristotle, being a rational animal is not just being an animal *and* being rational, but rationality is one of the forms that being an animal can take. Being a rational animal is the realization of all the powers of being an animal in a specific and distinct way. This means that there will be kinds of perception, volition, emotion which are specific to rational animals, and which are absent in nonrational animals, where those same powers will be developed in an essentially different way. So, while there will be some commonality between perception in rational animals and nonrational ones, the commonalities will be very generic. Rationality informs every other function, making it develop in a specific way. This can also be said about intuitions and emotions. Contrary to theories of moral cognition following the standard approach, which envision emotion and cognition in moral judgments as two encapsulated faculties, which cannot influence their respective inner processing, transformative theories hold that the emergence of rationality in humans changed the very essence of the emotional processes which are at the foundation of our moral judgments. In this chapter, I am going to review some theories that posit that intuitions are "educated" by the advent of rationality, creating a type of intuitive judgment markedly different from basic, unsophisticated, or automatic reactions. In the final section, I am going to present my ideas for improving existing theories of the education of intuition.

## 4.2. THE EDUCATED INTUITIONS MODEL

In this section, we are going to consider a theory of the education of intuitions whose aim is to defend rationalism from the intuitionist challenge against moral rationalism, that, as we saw in the first chapter, consists in the claim that, as moral judgments are intuitive and moral reasoning is post-hoc, reasoning is confabulatory. The educated intuition theory proposed by Hanno Sauer (2012a) tries to show that reasoning can, at least sometimes, not be confabulatory, as being intuitive does not exclude being rational, that is being the product of a process of conscious deliberation: most moral judgments are rational habitual actions, behaviors that were done consciously but which have now automated their *modus operandi* without changing their rational, deliberate essence. When a particular judgment is made, it is almost always purely intuitive, but the intuition has been shaped, over the course of the subject's life, by her acts of private reasoning; and there are subconscious reasons behind most of our judgments. Habitualization has made these reasons migrate from consciousness to subconsciousness, but post-hoc reasoning can recover them; and we have means to tell

apart this genuine post-hoc reasoning from mere confabulation. Deliberation can be causally effective, and moral judgments are rational anyway, so the intuitionist challenge is dispelled.

One might defend rationalism from Haidt and colleagues' intuitionist criticism by showing that it is not empirically true that most moral judgments are intuitively made or that, under most circumstances, moral reasoning is a post-hoc endeavor; but Sauer aims at defending rationalism while conceding to intuitionists that these ideas are correct. He argues instead that the problem stands in the way in which Haidt defines intuition and deliberation, definitions which imply that if a judgment is the product of intuition, it cannot have been influenced by reasoning processes (the *incompatibility thesis*). If incompatibility is false, the evidence collected by Haidt and colleagues in favor of the pervasiveness of intuition in moral judgment and of the post-hoc nature of moral reasoning would not lead to the view that moral reasoning is a causally ineffective confabulation.

In a nutshell, Sauer claims that most moral judgments are the result of "educated" intuitions, falling within the category of *habitual actions*. Educated intuitions present the *modus operandi* of intuition, while at the same time preserving the functional characteristics allowing us to properly consider them rational. These intuitions are the result of a migration of the processes of moral judgment from consciousness to automaticity (the *habitualization* of judgment). Habitualization preserves the rational character of a process. So, while each judgment feels purely intuitive, the intuition itself is a sophisticated, goal-directed, and rational process. Habits are defined to be automatic and intuitive kinds of behavior. Most of our daily actions are habitual. Yet it can be argued that they are rational. Often there are subconscious reasons for our daily actions. Those reasons are not present in consciousness when we act, but, if the need arises, we can retrieve them later with our reasoning; and there is a series of methods which allow us to tell apart sincere retrieval of reasons from mere confabulation (Sauer, 2012a, pp. 264-265). We can interpret the defense of rationalism by Sauer as an attempt to overcome the standard approach: cognition influences the inner workings of intuitions, so that educated intuitions are neither fully cognitive nor fully emotional. While in the moment in which the judgment is made it is felt as affective and intuitive, the judgment can still be said to be rational, and we can articulate sensible reasons for it, which are the actual reasons that were subconsciously present at the time of actuation. Rationality has nothing to do with the fact that a judgment or choice is the result of an effortful mental process, or that the reasons for the action are consciously accessible to the subject.

> PARITY PRINCIPLE FOR AUTOMATIC PROCESSES: "If we would not hesitate, on the basis of its functional characteristics, to call a process

> 'rational' were it performed consciously and effortfully, but, as it happens, it has become habitual and automatic over time, then we should not hesitate to call it 'rational'." (Sauer, 2012a, p. 261)

At least some of the time, a habit is an action which we previously performed effortfully and consciously and that we have automatized. The action has changed its *modus operandi*, but the same reasons for the action are operative, just they have moved from being conscious to subconscious. Sauer thinks that the difference between rational and irrational habits is to be found in whether the habit, while automatic, is done for subconscious reasons, reasons which might have been conscious prior to the process of habitualization.

To know whether habits can be considered rational we have to ask what distinguishes full-blown actions from mere behavior. According to Pollard (2005), philosophers' usual answer appealed to the *reasons theory of rational action*, the idea that full-blown actions are done for reasons. Theories of rational action must satisfy the *conceptual constraint*, the idea that most actions made by rational creatures are rational actions, and the *distinctiveness constraint*, the idea that there must be criteria that clearly distinguish rational actions from the behavior of non-rational creatures. The reasons theory of rational action meets the distinctiveness constraint, as we can suppose that only we humans, but not other animals, can act for reasons; but it fails to meet the conceptual constraint as, Pollard argues, most actions made by humans are habitual actions, which cannot be said to be done for reasons.

When can an action be said to be done for reasons? One possible requirement is the *conscious accessibility* of reasons. It is easy to see why this requirement would be too strict. The reasons for a given action are most often absent from our consciousness at the time when we act. A theory of rational action accepting the conscious accessibility requirement would thus fail to satisfy the conceptual constraint. An alternative is that a reason counts as the reason for an action if the agent is able, even after the action, to explain why she acted in such a way by citing that reason. While the reason was not consciously present at the time of actuation, it was subconsciously present, as the agent can later utilize it to explain herself. Call this the *subconscious accessibility requirement*. While Sauer seems to accept the need for subconscious accessibility, Pollard finds two problems with the requirement: first, for most habitual actions we cannot find much post-hoc justification (again, if we say that these actions are not rational, we would violate the conceptual constraint); second, the requirement does not gives us means to distinguish between genuine post-hoc explanations (those which allow us to recover our subconscious reasons) from confabulatory rationalizations. We don't really know the motives behind most of our actions, and we just make up what we believe is the most sensible explanation, just like we do when we try to explain the actions of others.

Pollard's aim is to explain how our habitual actions can be rational. Habits constitute a large part of our daily actions. If we want to produce a theory of rational action which meets the conceptual constraint, we should explain how they can be rational. Pollard's solution is that we should abandon the reasons theory of rationality and adopt a *permissive conception* of rationality. A permissive concept cannot be defined positively, rather it is defined in relation to some standard it does not meet. An example is that of legal actions. There is nothing common between all actions that are legal except the fact that they do not violate the law. So, there is no positive feature shared by all legal actions. Rather, legal action is defined in terms of what is not illegal. Conceiving rationality as a permissive concept leads us to define rational actions in terms of all what is not irrational. We have, Pollard argues, a much better grasp, both scientific and intuitive, of what constitutes an irrational action than of what is shared among all rational actions. Adopting a permissive view of rationality also allows us to conceive of habits as rational, for the simple reason that they are not usually taken to possess the features which make an action irrational.

Adopting Pollard's view, rationality has nothing to do with the fact that an action is the result of an effortful mental process, or that the reasons for the action are consciously or even subconsciously accessible to the subject. So, we might conclude that a process can be fast and effortless and, at the same time, also be rational. How can we be sure that an action is rational if conscious accessibility is not met? As we have seen, Pollard (2005) shows us two alternatives: either an action is rational because it is done for subconscious reasons, or an action is rational even without conscious accessibility, just because its functional properties are such that we would not claim that the action is irrational. Sauer (2012a, 263-264) prefers to adopt the subconscious accessibility requirement rather than the permissive theory of rationality. Contrary to Pollard, Sauer believes that habitual actions are done for reasons, and this explains why there is some structure to the way in which we acquire and display habits. Habitual actions serve the agent overall goals, and they are motivated by said goals.

According to Sauer, educated intuitions are rational insofar as they are done for subconscious reasons, and because they result from a process of automatization from previously effortful and conscious activity. Sauer can thus argue that most moral intuitions are based on reasons as they are informed by the subject's reasoning processes. While it is true that most moral judgments, taken as individual acts, are produced without the involvement of conscious consideration for reasons, conscious evaluation of one's judgments and the reasons for them can educate intuitions themselves *over time*. In other words, it is not true that conscious reasoning can only be effective in moral judgment by allowing the subject to consciously evaluate reasons in a process of deliberation, but it can

also be effective by modifying (educating) the unconscious processes which give rise to her moral judgments.

It is worth highlighting how this defense of rationalism requires a change of perspective from a *synchronic* view of moral judgment – one where the cause of each judgment is traced back to the immediately preceding psychological event – to a *diachronic* view – one where the long-term effects of reasoning in shaping ("educating") intuitions are considered. Synchronically, each moral judgment is, most of the time, done intuitively, without much concern for reasons. Diachronically, the intuitions which produced the judgment were shaped by the subject's acts of private deliberation.

### 4.3. INTEGRATING INTUITION AND DELIBERATION

The same idea that processes behind moral judgment can present the *modus operandi* of intuition and the functional properties of rationality has been used by Jillian Craigie (2011) against the strict division of labor envisioned by the Central Tension Principle. In her model, cognitive reflection and emotional intuition still play different roles, although they are integrated roles rather than mutually exclusive ones. There is no particular kind of response which is primarily associated with emotion rather than cognition ("there will be a less rigid association between particular normative frameworks and particular cognitive processes than Greene proposes", Craigie, 2011, p. 62), but cognitive processes, which are slower, can take into consideration features of the situation which the emotional systems cannot. The progressive automatization of cognitive processes as well as the progressive metacognitive education of intuitions will lead to more refined kinds of intuitive moral appraisal of a given situation. The interesting empirical question within Craigie's theory is no longer about which kinds of response are primarily associated with emotion vs. cognition; but rather about how *moral skill* is formed. This skill, which includes both cognitive and emotional aspects, consists in the educated intuitions allowing the moral agent to efficiently evaluate a given moral problem. The emotional part of these skills will provide the raw motivational force bringing the subject to assign a moral weight to certain features of a situation, while the cognitive part will consider further aspects of a situation and will *over time* lead to more sophisticated intuitive responses.

Craigie argues that Greene's attribution of a specific kind of judgment (deontological judgments) to a specific kind of processing (affective processing) overlooks the fact that acts of reasoning over intuitive outputs modify the intuitive outputs themselves, educating them. Here we find the same basic ideas of Sauer: that each moral judgment is the result of refined intuition which present both cognitive and affective aspects, that processes of moral evaluation migrate over time from the conscious type of processing to the automatic one, and that the focus of traditional theories on a narrow timespan view is the source of their

inability to properly take into consideration the effects of reasoning on intuitions; although these elements are not aimed at defending the causal effectiveness of reasoning but rather at showing how deliberation and intuition cannot each be linked to a specific kind of mortal response.

Craigie's (2011, p. 53-54) main argument is that "moral cognition is best understood as a species of Kahneman and Frederick's dual-process model of decision making" which show that there are strong reasons to "resist endorsing simple sentimentalism or rejecting rationalism" and which conceives "emotional intuition-generating processes and reflective processes as operating in an integrated way in moral decision making, assigning metacognition an essential role in the monitoring and shaping of moral intuitions". This means that while most of our moral judgments are produced by affective intuitions, we are still legitimate to say that they are rationally controlled, as metacognitive processes make them consistent with moral principles and norms of rationality. Indeed, acts of metacognition shape, over time, our intuitive emotional reactions, making them sensitive to features we previously were oblivious to. In this sense, we can still draw a distinction between processes which generate emotional intuitions and processes of cognitive deliberation, but the two, rather than being linked with different kinds of responses, as Greene proposes, are *deeply integrated* in the production of ordinary moral judgments.

Kahneman and Frederick's (2002) dual-process framework was motivated by the observation that, when faced with a difficult question, people tend to subconsciously substitute the judgment about the difficult property with a judgment about a more readily accessible property. Emotion often functions as a heuristic, and the kind of computation which it performs is to be contrasted with more cognitive approaches. So far, this model is highly compatible with Greene's idea of a division of labor between emotion and cognition: emotion focuses on features (like the use of personal force) which are different from the kind of features cognitive deliberation would focus on. However, in Kahneman and Frederick's account there is also the possibility of an integrated functioning of cognition and emotion, and, more generally, of deliberative and intuitive processes. Consider perceptual judgments. We adults automatically take the blurriness of an object to judge how far it is. In other words, we employ blurriness (an easily accessible feature) as a heuristic for distance (a feature which is harder to compute) – subconsciously applying a substitution in our judgment. However, this heuristic leads to a biased form of judgment because blurriness can be caused by things other than distance – like hazy whether. Some people are aware that haze makes things seem further away than they are, and they will (consciously at first) consider the information about the weather when making a judgment about distance. Craigie (p. 59, italics in the original) writes:

> "If the observer does take the haziness into consideration the resulting judgment will be something like: *the building seems a long way away but it's probably closer than it seems because of the haze*. In this case a System 1 process uses the blurriness of an object to generate an impression about its distance. This kind of process is fast and doesn't deplete valuable cognitive resources but it lacks the flexibility required to take into account factors that can distort impressions. On the other hand System 2 processes are metacognitive in nature and so are able to incorporate this kind of information in the decision making process. Because of the strengths and weaknesses of each system, competent decision making requires the integrated functioning of the two."

In the case of a distance judgment, to arrive at the correct conclusion one needs to employ both an intuitive process (the blurriness heuristic, necessary to have some way of evaluating distance) and a reflective process (the evaluation of the weather, allowing one to judge the limits of the blurriness heuristic. It is not like the reflective system can arrive, without the intuitive process, to a conclusion about distance. What is happening is that the reflective system is *judging over* the intuitive input to arrive at a better judgment. So, the two systems are working together rather than independently one of the other. The reflective system is metacognitively evaluating the heuristic employed by the intuitive system, comparing it with information inaccessible to the intuitive system. This integrated functioning of the two systems can, over time, modify the intuitive input itself. The prolonged influence of cognition over emotion can modify the kind of intuitive judgment which is produced by emotion itself.

Craigie notes that this creates a tension between Greene's dual-process proposal and Kahneman and Frederick's overall dual-process framework, despite Greene and colleagues' (2004) belief that their theory falls within Kahneman and Frederick's more general model of cognition. Craigie (2011, p. 60) writes that: "Kahneman and Frederick also propose that over time integrated functioning of the two systems can result in the modification of intuitive responses such that they come to reflect, at least to some degree, the output of metacognitive processes. So although System 1 operations are proposed to be more evolutionarily ancient than System 2 processes, their outputs need not be considered less sophisticated." This is due to the fact that, in normal development: "complex cognitive operations eventually migrate from System 2 to System 1 as proficiency and skill are acquired" (Kahneman & Frederick, 2002, p. 51). Craigie holds that intuitive-emotional responses are not fixed products of our phylogenetic history and that they show cultural and individual variability.

In reviewing the model, Maiese (2014, p. 810) proposes that the emotional system is a nuanced response resulting from individual learning and socialization:

> "Very few theorists would deny that natural selection has retained certain neural structures based on their ability to generate adaptive behaviors given certain stimuli (Hanoch, 2005, p. 140). As a result of connections forged between certain forms of emotional arousal and certain forms of stimuli over the course of our evolutionary history, the experience of particular emotions serves as an alert system. However, like, I find it plausible to suppose that it is not only evolution, but also learning and socialization, which forge connections between stimuli and patterns of thought and behavior. As de Sousa (1987) points out, the 'paradigm scenarios' that define an individual's emotional repertoire are a matter of learning and development. The process of establishing links between feelings and certain characteristic objects is ongoing, so that emotions are largely a product of cultural influence and habituation rather than just primitive instinctive responses (de Sousa, 1987, p. 182). Emotion comes into play in moral cognition as a worldview and background orientation, built up over time, through which individuals interpret their surroundings."

Becoming an expert moral agent involves improving one's intuitions by means of prolonged metacognitive reflection over them. The resulting intuitions cannot be classified purely as emotional or cognitive: they have surely an emotional component, but also a strong cognitive aspect which is imbued into them by means of the previous acts of metacognition. They are "educated" intuitions. Conceiving the intuitive reaction to a moral problem as a nuanced reaction, responsive to moral principles, we wouldn't so readily adopt the normative conclusions of Greene's theory. To come to the normative conclusion that we ought not trust our deontological judgment, the dual-process theory has to assume that the emotional reactions giving rise to deontological judgments are less sophisticated than the output of the cognitive system. If our emotional reaction to moral dilemmas is informed by our previous acts of metacognition, there would be no reason to hold that the response coming from such reaction is less trustworthy than a more time-consuming consequentialist analysis of the dilemma. Even if deontology is really associated with emotion and intuition, insofar as that emotion is progressively educated by our cognitive abilities, we would have no reason to discount the output of that emotion and no reason to only trust our explicit cognitive responses. Craigie (2011, p. 60) writes:

> "[…] an important difference between these approaches is that Kahneman and Frederick's model emphasizes the integrated functioning of the two systems and the subsequent modification of intuitive responses, while Greene's proposal characterizes the two types of process as being fundamentally in competition with one another. In addition, Greene (2008) characterizes intuitive (System 1) responses in the context of moral decision making as relatively fixed products of our evolutionary history. It is suggested that these moral intuitions can be 'shaped and refined by culture bound experience' (Greene et al., 2004, p. 398). However, there is no further discussion of the extent to which moral intuitions are amenable to modification, particularly modification as the result of reflection. The proposal's clear focus on the idea that moral judgments are driven by either one system or the other, and the need for reflective processes to override intuitive responses, suggests that the scope for reflective modification of moral intuitions is assumed to be minimal."

According to Craigie, emotions involved in moral decision-making are sophisticated responses, which show cultural variation and individual flexibility. Emotions provide the default response ("most people, most of the time, base their moral appraisals on intuitive responses associated with affect", p. 63). However, this response can be consistent with moral principles and reflectively endorsed beliefs ("an intuitive judgment is controlled by a System 2 process [the cognitive system] as long as the intuition would be corrected in the event that it diverged from the relevant reflectively endorsed response", p. 68). The intuitive/emotional response also becomes progressively more sophisticated over a lifetime of training as a moral agent (the effect of cognitive reflection on intuition consists in a "transformation [of intuition] over time rather than the ability to revise or reject particular intuitive responses, p. 66). The competent moral agent can trust his emotions because of the lifelong training he has had as a moral decision-maker. One might think that this claim is consistent with Greene's idea that we can trust moral emotions in familiar contexts. I think, however, that Craigie's theory goes one step further. It is not just that emotion is an automatic response that is apt to tackle specific problems (familiar problems), rather emotion is a flexible, nuanced and rational kind of response insofar as, in our lives, emotion and cognition are integrated in providing a response to moral problems.

### 4.4. THE DIACHRONIC INFLUENCE OF DELIBERATION OVER INTUITION

Social intuitionism and the dual-process theory of moral decision-making have overlooked the effects of reasoning over moral intuition because they focused on a narrow

timespan view of judgments, whereby each of them finds its cause in the immediately preceding psychological event; events which are almost always intuitive and emotional. The claim of social intuitionism that all that is needed to explain the formation of a moral judgment is the preceding flash of emotion relies on the assumption that moral agency can be properly characterized as a stream of isolated moral verdicts, each of which is explained by the psychological event that immediately precedes it (Sauer 2017, p. 11). Haidt and his colleagues then attempt to show that these psychological events are predominantly emotional in nature and that they suffice to explain moral judgments. Sauer argues however that social intuitionism rests on a theoretically problematic starting assumption: rather than seeing the moral agency of the subject as a set of disconnected and immediate moral judgments, there might be reasons to consider each single judgment as stemming from the moral agency of the subject, a moral agency which can only be appreciated by looking at the person as a psychological entirety and as a temporally extended moral subject. According to Sauer, morality is not a bundle of disconnected judgments, but the result of a comprehensive and integrated process of personal and cultural development. Focusing just on the immediate proximal causes of moral judgments makes social intuitionism blind to the role that acts of reasoning, even when they are performed individually, may have over our judgments. One can in fact concede that most moral judgments are automatic, and even that their immediate proximal cause is mostly emotional in nature, but, if we extend our view of the subject, we may wonder whether those intuitions and emotions are themselves shaped by reasoning processes.

How to explain, on the other hand, the finding that emotion and cognition seem to produce two distinct kinds of responses when we face impersonal moral dilemmas vs. personal ones? First, there is a methodological worry which might make Greene's theory oblivious to the influence of reflection and metacognition on emotional intuitions. Greene and colleagues have collected their brain-imaging and response time data "at a single time point" (Craigie, 2011, p. 61). Whatever effect deliberation on one's past responses might have on one's current intuition can only be appreciated after repeated exposure to similar moral problems, while the focus of Greene and colleagues was on the immediate reaction of subjects to specific problems. This was also exacerbated by the unfamiliar nature of problems like the trolley and footbridge dilemmas. It is conceivable that the intuitions we have when facing more familiar problems present a greater influence from previous instances of deliberation. For Craigie, the association that Greene and colleagues have found between consequentialist judgments and cognitive deliberation depends on the nature of the dilemmas administered to the subjects. Craigie (p. 61-62, italics in the original) writes:

> "All the dilemmas, like the footbridge case, presented participants with a choice between a judgment associated with an emotional intuition (e.g., it's inappropriate to push the man) and a judgment consistent with a utilitarian principle (e.g., it's appropriate to push the man). There seems to be no reason to conclude that the reflective processes evidenced in the studies are associated *only* with utilitarian judgments since the anti-intuitive judgments in these studies were always utilitarian. […] examples in which intuitive responses can conflict with judgments based on other moral principles include the right for same sex couples to get married or have children. These issues expose a tension between intuitive disapprobation (at least among social conservatives) and considerations regarding equality or the right to pursue happiness, rather than utilitarian concerns."

Greene might rely on a skewed set of problems, those where the intuitive conclusions are contrasted with consequentialist considerations, but in other problems, Craigie suggests, intuition and emotion might be in contrast with other kinds of considerations – and even deontological principles about rights and duties.

Just like Sauer and Craigie, Campbell and Kumar (2012) also believe that the reason why Haidt concluded that moral reasoning is confabulatory is due to limitations in the methodology with which he conducted experiments. The reason is to be found in the fact that much of the experimental research tends to be "temporally limited", that is, focused on the proximal causes of a single judgment, while often "the effect of reasoning on moral judgment emerges only over a significant length of time" (Campbell and Kumar, 2012, pp. 286-287). A parallel of the diachronic effects of reasoning can be found in perceptual judgments. In *The Modularity of the Mind* (1983), Jerry Fodor proposes a distinction of psychological functions along lines which have now become common to distinguish the two types of processing, telling apart perceptual systems from central processing systems. Fodor argues that perceptual systems are cognitively impenetrable by central processing systems. One example of this can be found in the fact that visual illusions persist even when the subject knows they are illusions. However, as reported by Campbell and Kumar (2012), further studies have shown how there are cultural differences in how people perceive visual illusions: the Müller-Lyer illusion is dependent on whether the subject has lived in a "carpentered" environment – that is an environment containing walls and buildings and other objects with sharp right angles. Those living in uncarpentered environments seem to be immune to the Müller-Lyer illusion. While the single act of reflection cannot change how one perceives the illusion, having lived all of one's life in a certain type of environment has effects on how one perceives the illusion. In the words of Cambell and Kumar (p. 282), this

shows how "the visual system is penetrable – not synchronically but diachronically". They then ask if a similar phenomenon can be found in moral judgment. If moral intuitions can change over time, given the penetrability from acts of reasoning. This should explain not only the individual growth in one's intuitive moral reactions (becoming adult, we have more refined intuitive moral reactions), but also moral change within societies, whereby things that were universally considered right (like slavery) become universally considered wrong, and vice versa.

Campbell and Kumar (2012) focus on a particular mechanism by which reasoning can diachronically penetrate moral intuitions: *moral consistency reasoning*. It is undesirable for subjects to have inconsistent moral beliefs: if there are principles making an action morally right in one circumstance, then subjects will strive to consider that action morally right in all morally similar circumstances. When someone points out an inconsistency in the moral beliefs of a subject, the subject will feel the pressure to find morally relevant differences between the two cases or to reject one of her prior moral beliefs. The authors make the example of Jan Baalsrud, a Norwegian resistance fighter pursued by the Nazis, who took refuge into a home in a small village. The family there had the choice of turning him over to the Nazis or helping him recover and escape, with the risk of being discovered and the whole village put to death in retaliation. Marius wanted to help Jan while his mother was concerned about the fate of her community. Marius won the argument by pointing out an inconsistency in her mother's moral beliefs: he asked her what a family in Oslo ought to do if it was him who was in Jan's shoes. Marius's mother had the strong intuition that they should turn Jan over to the Nazis to save her village, but she had the even stronger intuition that if it was Marius to be in Jan's position, the other family should have made him recover and escape. The two intuitions were clearly in conflict with one another, and there was no relevant moral difference between the real situation they found themselves in and Marius's imagined situation. So, Marius's mother had the choice to completely undermine her credibility as a moral agent, arguing that it was right to save Marius but not Jan just because Marius was her son, or to reject the weakest of the two intuitions and accept to host Jan for the time it was needed. She opted for this latter option and, fortunately, Jan recovered and was able to escape without alerting the Nazis. As we can see from this example, subjects strive to maintain consistency between their moral intuitions and, when inconstancies are pointed out by others, they must adjust their intuitions either by rejecting one intuition (like Marius's mother did) or by finding a relevant difference between scenarios which would justify a difference in moral choices.

Campbell and Kumar (p. 287) argue that moral consistency reasoning can shape intuitions, such that future moral intuitions resent its effect even when new acts of reasoning

are not performed. Moral consistency reasoning can be a means by which reasoning diachronically penetrates moral intuitions, educating them, just like having lived in a carpentered environment changes visual intuitions. As an example, consider the change of attitude of straight people towards homosexuality. During adolescence, many think that gay sex is disgusting (an emotional reaction) and wrong (a reasoned belief, coming from the influence of parents, peers, and the religious authority). Then, many reject their belief because they come to doubt the religious condemnation of homosexuality and because they become aware of the immense variety of sexual practices in nature. Still, the emotional reaction to gay sex remains. But people can think over their reaction and see if this moral feeling is consistent with the feeling they have in similar circumstances. What difference between gay sex and straight sex would justify reacting with disgust in one case and not the other? People ideas vary, but many cannot find any strong morally relevant difference, and thus they come to believe that homosexuality is morally okay, and they lose most of their emotional response against gay sex. Moral consistency reasoning has made their affective intuition change, in such a way that the new intuitions reflect the previous acts of metacognition. Similar processes, but a collective scale, are responsible for historical moral change within society, whereby certain previously marginalized groups (homosexuals, women) acquire rights and certain practices (like slavery) become illegal and stigmatized. The effect of consistency reasoning on moral intuitions shows that the social intuitionist belief that moral reasoning is just a rationalization is false. Intuitions are refined by reasoning, such that previous acts of reasoning, and of moral consistency reasoning, "educate" intuitions.

### 4.5. AFFECTIVE FRAMING

The models reviewed in the previous sections aim their critical force against different targets, but they are united in the same basic ideas. The kind of functional relation they envision for cognition and emotion in moral judgment goes beyond the standard approach: it is not just that cognition and emotion can interact, but this interaction changes their inner nature giving rise to processes (educated intuitions, habits, skills) which can no longer be analyzed through the lens of the categories of cognition and emotion. There is no specific kind of judgment which is attributed to emotion rather than cognition, and there is no specific kind of role that emotion does, and reason does not. There is more: phenomena like moral consistency reasoning are not just an effect of reasoning over intuitions, but this effect is modulated by emotions themselves. In other words, as one line of argument against the standard approach looks at the influence of reasoning over emotions, a complementary line of argument looks at the influence of emotions over cognition.

The phenomenon of affective framing leads us to conclusions similar to those made by Craigie and Sauer about the education of intuitions, but with a different spirit: while the education of intuitions makes us consider the effect that cognition has on emotions (metacognitively educating them over time), affective framing makes us appreciate the effect that emotions have on cognition (serving as a necessary precondition to cognitive evaluations). Maiese (2014, 813-814) argues that all moral judgments, including those following from consequentialist considerations, are strongly influenced by emotions. Maiese also argues that emotion is *necessary* for moral judgment. Emotion is a necessary precondition because it allows us to focus our attention on only those features that can be relevant for the judgment. Maiese (pp. 814-815, italics in the original) writes:

> "[…] in the context of decision-making, it appears that the rational cognition involved in cost-benefit analysis is insufficient to limit the amount of information that one takes into account in order to arrive at a reasonable decision in a timely manner. To avoid sifting through the exponentially large range of potential outcomes and overcome this problem of potential information overload, agents must focus their attention on those features of the world that strike them as most relevant and important given the context and the nature of the decision being made […]. Deliberation then proceeds on the basis of a restricted set of features […]. Likewise, in the context of moral assessment, the meaning that an observer assigns to someone else's action depends on the inferences and interpretations this observer makes about the intentions and circumstances of the actor. Any particular action or behavior can have a range of different interpretations depending on which aspects of it the observer highlights as significant. […] the starting points for reasoning are selected via a *framing* process."

The mechanisms that enable framing must work quickly, before theoretical reasoning is started, and they have an essentially evaluative character, reflecting the judging subject's background and interests. In a nutshell, the framing which is a necessary precondition for moral judgment is realized by an affective system, thus emotions are necessary for moral judgment ("The way in which an agent interprets and affectively appraises a situation, highlighting certain factors while ignoring others, not only *crucially influences*, but also *enables* his or her processes of deliberation", p. 816).

This idea is compatible with the view that emotion precedes cognition. First, we identify, through a system of affective framing, those features of a situation which have moral relevance, then we compute, through cognitive reflection, the judgment we hold to the situation. Such a proposal would be akin to the currency-like emotions Greene believes

to be involved in consequentialist judgment. Crucially, this view is not shared by Maiese. She believes instead that "affect processing occur[s] *during* moral evaluation" (p. 816, italics in the original). The reason is that the development of complex cognitive evaluations is intertwined to the development of a progressively more sophisticated system of emotional attunement, in what can be described as a psychological feedback loop. Maiese writes (p. 816):

> "The claim that emotion and cognition normally are deeply integrated, rather than operating separately, is supported by the fact that the development of one's cognitive capacities ordinarily occurs together with the development of ever-sophisticated patterns of emotional attunement. As we begin to cultivate certain emotions in ourselves, we may develop a capacity to grasp the significance of certain events and become attuned to nuances we could not previously recognize. In addition, there is good reason to think that the perceptions and interpretations which comprise one's sense-making activities always are partly a matter of what one desires, values, and deems significant."

This proposal can easily be made compatible with Craigie's model, by considering how Craigie is focusing on the effect that cognition has on emotion (acts of metacognition educate emotional responses), while Maiese is focusing on the effect that emotion has on cognition (more refined emotional responses allow to identify new, previously overlooked, features of the situation as morally relevant and thus they enable more sophisticated kinds of reasoning). Saunders (2016) proposes that the "ever-sophisticated patterns of emotional attunement" Maiese refers to find a prerequisite in the development of cognitive skills.

The remarks by Saunders highlight a crucial feature of Maiese's theory. While she insists on the effects of emotion over moral judgments, her model is not a sentimentalist one. Rather, her proposal is aimed at showing how cognition and emotion interact, to an even greater degree than that proposed by Craigie. Maiese writes (p. 818):

> "[…] then the usual distinction that dual-process theorists make between deliberation and conscious reasoning on the one hand, and emotion on the other hand, may be far blurrier than they thought. Reasoning processes crucially involve, and in fact depend upon, an immediate, affective, pre-reflective, non-inferential interpretation of our surroundings of which we often are not at all self-reflectively aware. Ultimately, this often leads toward further deliberation about possible courses of action or examining the reasons for thinking an individual or group has behaved rightly or

> wrongly. I certainly do not deny that reasoning, deliberation, and the application of principles play a crucial role in decision-making and moral judgment, but instead want to propose that affective framing is bound up with these more reflective processes, and in fact plays a strictly necessary and integral role in helping them to get off the ground."

Maiese (p. 824) explicitly disagrees with authors like Prinz (2006), who believe that the significant influence that emotion has on moral judgment (emotion being necessary to moral judgment) implies that moral judgments are nothing but the expression of one's emotions.

The emphasis Maiese puts on the necessary influence of emotion over cognition finds a complement in Campbell and Kumar (2012) stressing that we should not conceive moral consistency reasoning as "the application of our general capacity for reasoning in system 2" (289), rather consistency reasoning is both reasoned and affective. One should contrast moral consistency reasoning with "principle reasoning", the application of a certain moral principle (whether consequentialist or utilitarian) to certain empirical facts to infer a moral conclusion. According to Campbell and Kumar, principle reasoning takes place entirely in the cognitive system: it is conscious, controlled, and effortful (p. 291). On the contrary, while moral consistency reasoning surely involves a certain amount of reasoning (describing the different circumstances and trying to find relevant differences), the effect that those circumstances have on us and the choice of what we believe to be a morally relevant difference are almost entirely dependent on our emotional system. Marius's mother *feels* that it would be right to turn over Jan to the Nazis to save her village and she *feels* (more strongly) that it would be wrong if a family in Oslo did turn over Marius to the Nazis. Furthermore, she *feels* that there is no relevant moral difference between Jan's case and Marius's imagined scenario. The evaluation of the two scenarios is emotional, then considerations about the consistency of those emotional reactions are produced by reasoning. As an example of the integration of reason and emotion in moral consistency reasoning, Campbell and Kumar (2012, 292-294) offer an analysis of Singer's debunking argument. The argument is based on moral consistency reasoning: first we have an affective intuition towards the drowning child case (we should help him!) and we do not feel a corresponding intuition towards the faraway starving child. Our cognitive system recognizes that the two responses are opposed and starts searching for differences in the two scenarios which would motivate a difference in our judgment. Let us say that our reasoning finds distance as a possible morally relevant difference between the two scenarios. This response of the cognitive system is fed back to the emotional system. The emotional system might provide the intuition that distance is not a morally relevant difference. Then, given that our reasoning cannot identify any other morally relevant difference, we feel the

need to abandon the weakest of our two starting intuitions, that is the intuition that we don't have any particular moral obligation towards the starving child. We feel in a certain way in the case of the drowning child and we feel in a certain way in the case of the starving child, we reason that these feelings are incoherent, and we feel that there is no relevant moral difference between those cases (in the sense that we cannot attribute an emotional weight to the differences identified by reasoning). While reasoning guides the process by noticing inconstancies in our feelings and trying to find motivations for them, emotions function as necessary inputs at every step: in the evaluation of the two scenarios and in the evaluation of their differences. In this sense, moral consistency reasoning is a primary example of the effect that reasoning can have on emotions (recall that through consistency reasoning people change over time their emotional attitude towards homosexuality) and of the modulation that emotion itself has on this effect (serving as input to moral consistency reasoning).

# 5. DEFENDING MORAL INTUITIONS

"It is important to learn to be surprised by simple facts."

- *Noam Chomsky*

It is a simple fact that most of us can understand and produce sentences in our native language with ease and speed, even though many of us may have some trouble making explicit the rules of syntax we are nonetheless intuitively applying. Our linguistic intuition is mostly trustworthy and effective, showing how sometimes intuition can lead us to master complex domains in a reliable way. One of the simple facts a theory of linguistics should explain is how people possess such a reliable linguistic intuition. Other times, intuition has no value for science. We all have naïve intuitions about physics that lead to mistakes and misconceptions about the physical world. If science took our physical intuition as a source of knowledge, it would be as limited as Aristotelian physics. In this chapter I am going to argue that moral intuitions are more like those about language than those about physics. The main consequence of conceiving moral decision-making as constituted by educated intuitions is that moral intuitions might be surprisingly nuanced and accurate. Just like skill acquisition in a field of expertise leads to successful attunement to that complex domain, so the acquisition of a moral skill leads to the successful intuitive attunement to the complexities of social life.

I am going to show how the direct route debunking argument put forward by Greene fails to show that deontological philosophy is worse than consequentialist philosophy because moral intuitions do not react to morally irrelevant factors. The mental experiments of Singer's Puzzle, Haidt's Twins Problem, and the trolley problem have flaws in their designs which give the impression that intuition is reacting to morally irrelevant factors – distance, the presence of incest, and personal violence respectively – while this is not the case. The problem stands in the highly artificial nature of the presented problems. Rather than blind automatic reactions related to the ancestral environment and set off by the presence of specific factors, moral intuitions react to the general narrative implicit in problems the subject faces – a method which is reliable in most of the circumstances in which intuition was trained, but which is set to fail in the aforementioned problems. Similarly, the indirect route fails to reach the conclusion that consequentialist philosophy is better because, as the skill model shows, the formation of intuitions is not a rigid process.

## 5.1. THE TWINS PROBLEM RECONSIDERED

Consider the trolley problem. Greene argues that the difference in judgment between the switch scenario and the footbridge one is the fact that the emotional system is elicited

by the presence of personal violence. One problem with this argument is that there seems to be cases where people tend to judge numerically identical dilemmas in different ways (deontologically vs. consequentialistically) even when no close-up violence is involved. Consider this dilemma by Hauser and colleagues (2007):

> HEAVY OBJECT DILEMMA
>
> Ned is walking near the train tracks when he notices a train approaching out of control. Up ahead on the track are 5 people. Ned is standing next to a switch, which he can throw to turn the train to an aside track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. The heavy object is 1 man, standing on the side track. Ned can throw the switch, preventing the train from killing the 5 people, but killing the 1 man. Or he can refrain from doing this, letting the 5 die. Is it morally permissible for Ned to throw the switch?

Here there is no close-up personal violence involved and yet subjects often provide the deontological response that it is not permissible to press the switch. To defend the normative conclusion of the evolutionary argument, one would need to find another morally irrelevant factor (besides personal violence) which would explain why we judge these cases deontologically when there is no personal violence arousing our emotions. Greene and his colleagues (2009) have proposed that such morally irrelevant factor is "personal force" defined (p. 365) as follows: "an agent applies personal force to another when the force that directly impacts the other is generated by the agent's muscles, as when one pushes another with one's hands or with a rigid object". This would allow us to explain the why we switch to a deontological judgment in such dilemmas, but it is hard to see why natural selection would have made our ancestors sensitive to the distinction between moral situations involving personal force vs. moral situations devoid of personal force. If the difference between cases where we judge deontologically and cases where we judge in a consequentialistically does not depend on factors which were shaped by our evolutionary history, then such difference does not depend on morally irrelevant factors. Personal force is not a viable candidate because, while it would account for the difference between our judgments on numerically identical scenarios, it does not also make justice of the evolutionary argument (Sauer, 2012b, pp. 797-798). In order to save the normative claim of Greene's theory from the objection which comes from the heavy object dilemma and similar scenarios, one would need to find a factor of a situation which would satisfy these two requirements: (1) allow us to explain why we switch to a deontological judgment when, were that factor be absent, we would judge the same situation consequentialistically; (2) be

a morally irrelevant factor, that is a factor which does not depend on the morally relevant features of the situation but on our specific constitution as beings shaped by natural selection. To my knowledge, Greene and his colleagues have not made a proposal which would satisfy both these requirements.

The problem coming from the heavy object dilemma illuminates a more general difficulty with the direct route argument: the idea that intuitions are "primitive" states which react to factors we are only sensitive to due to our phylogenesis seems false. The question then becomes what features of the trolley problem our intuitions are responding to. These features should explain our difference in judgment (deontological in the footbridge case, consequentialist in the switch case). My answer to this question is that, given the idiosyncratic nature of the education of intuitions in each individual, there are no specific factors to which intuitions apply to, but rather intuition responds to the general narrative implicit in the story, following the personal interpretation of the subject.

Consider again the Julie and Mark story administered by Haidt and colleagues. The usual intuition is that the act of the two siblings is morally wrong. Haidt argues that this intuition does not react to the actual reasons that would make the act problematic, as the story is told in such a way that most of the usual reasons don't apply. The conclusion of Haidt's argument is that intuition is reacting to incest *in virtue of* it being incest: intuition is blind to reasons and is a simple automatic reaction towards an act we have evolutionary reasons to be disgusted by. Railton (2014, pp. 847-851) however argues that even in Haidt's example intuition might be reacting to other features rather than incest itself. If intuition is indeed a complex educated intuition, it might reflect a worry about the psychological consequences for the siblings, it might evaluate how poorly their idea of making love was, and so on. Rather than a basic emotion targeting incest *in qua* incest for evolutionary reasons, intuition might reflect a series of complex states direct to the whole narrative sequence of the Julie and Mark story: indignation for their action, worry for their foolishness, and so on. Rather than being activated by the simple detection of a frowned upon act like incest, intuition might respond to the general story being told, in all its psychological and social complexity.

But how could intuition end up condemning the two siblings if the story itself says they experienced no negative consequences? Because one can, very rationally, condemn people for having put themselves or others in danger even when no actual danger has emerged. Railton makes the example of two people playing Russian roulette with a gun. Even if the two survive and somehow don't experience lasting psychological consequences, it is perfectly reasonable to condemn their action, as they have put their lives at risk. If intuition is really responding to the psychological and social subtleties of the Julie and Mark

story, it might reach a similar conclusion: despite the story telling that the siblings incurred no harm (not even psychological harm) and took some precautions against the more direct consequences (like disease or pregnancies) the two still put their emotional, psychological and social life at risk, and this is enough to condemn them in a perfectly rational manner. If Railton hypothesis is correct, when we hear the story of Julie and Mark our intuition reacts to this story of risk and foolishness and, having in an instant assessed that the siblings have put themselves in danger, it condemns them. In this scenario, intuition would be much deeper in its analysis of the situation than a simple gut reaction to a specific action, like incest.

### 5.2. THE TROLLEY PROBLEM RECONSIDERED

I think that similar considerations can be made about the trolley problem. The trolley problem is structured in such a way as to constrain our intuitions. Either we choose one option (hitting the switch/pushing the person) saving five lives at the price of one, or we choose the other (not acting) sacrificing the five lives. The structure of this problem is such that it removes all the complexity usually found in real-life situations. These are ways in which the trolley problem simplifies what we usually face:

1. It asks us to imagine that we are the only one that can intervene in a situation, while usually there are people with which we might confront to find a solution.
2. It puts us in front of two options giving us no other way of acting and excluding the possibility that some other solution might come to mind if we carefully analyze the situation, while in most real-world moral problems we have a huge variety of available actions, many of which are creative solutions.
3. It specifies that either of the two options has a specific amount of damage that would be caused or prevented, while in most real-world problems evaluating the consequences of an action is complex and there is a lot of uncertainty.
4. It assumes that the relation between actions and consequences is fixed and that we are aware of it beforehand, while in most real-world cases exogenous factors might intervene in between action and consequence (in the footbridge case, we are asked to exclude the possibility that, once on the tracks, the fat man wouldn't still be able to stop the train, but if we really found ourselves in such a scenario this doubt would surely come to mind).

My point here is that by making all these assumptions, the trolley problem might lead researchers to *underestimate* the powers of our faculty of intuition. When we are put in front of real-life problems where there are other people, multiple options, and a lot of uncertainty about consequences of actions, intuitions might function really well in guiding us, because this is the kind of scenarios they were trained to deal with. When instead we find ourselves

confronted with a weird mental experiment like the trolley problem, intuitions might be led astray *by means of their own strength*. We are confronted with either the switch or the footbridge scenario, and intuition reacts as it usually does in normal circumstances, responding *not* to morally irrelevant factors *nor* to specific factors in general, but rather to the overall *story* that is told. Intuition might wonder at why can't we ask for help by other people, it might ask if there is any means of alerting the workers on the tracks, it might doubt that we are the right kind of person to take care of such situation, it might envision new possible ways of action beside those proposed by the problem, it might question whether the weight of the person on the footbridge would be really enough to stop the train. In other words, if we are exposed to the story for the first time, intuition might treat this case as a usual real-life case, just as it is used to do with all other cases. Then, the experimenter recalls us that this is not a real-life case, but a *mental* experiment, and that all other considerations besides "would you hit the switch/push the person or not?" are irrelevant. Here, we look again at the problem, with our intuition caged by the strict requirements imposed by the highly hypothetical nature of the trolley problem, and intuition might then appeal to some lower-level feature. Maybe it is true that what *ultimately* explains our deontological turn is that intuition is repelled by the personal violence we have to exert on the person on the footbridge. But intuition only turned to this morally irrelevant factor because, by the way in which the experimenters had presented the problem, all other factors intuition could have considered were deemed irrelevant to the situation. What this means is that in real-life situations, intuition might be a good guide and indeed the only way to navigate the complexities of real-life moral problems.

### 5.3. SINGER'S PUZZLE RECONSIDERED

Similar considerations can be made in the case of Singer's Puzzle. A shared assumption of both those that consider distance a morally irrelevant factor (like Singer himself) and those who have tried to argue that it is actually morally relevant (like Kamm, 2007) is that the reason why we judge differently in the two cases of the puzzle is that our intuition is sensitive to distance. The problem is that in the original example made by Singer, distance is confounded by many other differences between the two scenarios. To solve this problem, Kamm (2007, p. 348) has designed this version of the problem which tries to minimize possible confounding factors:

> NEAR ALONE: I am walking past a pond in a foreign country that I am visiting. I alone see many children drowning in it, and I alone can save one of them. To save the one, I must put the $500 I have in my pocket into a

> machine that then triggers (via electric current) rescue machinery that will certainly scoop him out.
>
> FAR ALONE: I alone know that in a distant part of a foreign country that I am visiting, many children are drowning, and I alone can save one of them. To save the one, I must put the $500 I have in my pocket into a machine that then triggers (via electric current) rescue machinery that will certainly scoop him out.

These two cases contain the same number of victims whose suffering is equally serious. The agent cannot save all the victims, eliminating the possibility that our sense of moral duty in the near case depends on the fact that we are managing the "whole problem". Costs for the agent, means of helping, and probability of success are exactly the same in both scenarios. This minimizes the possibility that a difference in our intuition between the two cases depends on factors are than distance.

However, Kamm herself notes that in the Near Alone and Far Alone cases distance is still confounded with informational directness: in Near Alone the agent directly sees the drowning children with her own eyes, while in Far Alone the agent receives the information about them via some mediating mechanism. Nagel and Waldman (2013) presented subjects with scenarios where the two factors (distance and informational directedness) were deconfounded. Each participant was randomly assigned to one of four cases: far/direct (the agent uses binoculars to see drowning children that are 10km away), far/mediated (the agent watches a cell phone video of children drowning 10km away), near/direct (the agent sees children drowning nearby, but he can only activate a machine to rescue one of them due to a fence being in the way), and near/mediated (the fence is a wall that obstructs the view and the agent knows about the children thanks to a video). Each participant was then asked to rank how strongly she felt the need spend money to activate a machine that would save one of the children in the scenario she was presented with. The results showed that distance had no main effect on the participants' sense of obligation, as people felt equally obligated regardless of their distance to the victim. A major effect was instead played by directedness, as participants felt greater obligation in the conditions they were asked to imagine seeing the victims with their own eyes rather than being informed about them through a mediating mechanism. At constant levels of directedness, distance ceases to be morally relevant. Another factor which is usually confounded with distance is group membership of agent and victim. To deconfound the two, Nagel and Waldman (2013) presented to each participant with one of four cases. In the near/same condition the subject can save a person of his own town by donating blood (the person would otherwise die), in the far/same condition the person to be helped come from the same town of the subject but the subject is

momentarily residing far away, in the near/different and far/different cases the person to be saved is of another nationality. The results showed that participants felt a stronger moral obligation to help victims of their own community than those coming from a different country, regardless of distance.

Nagel and Waldman (2013) also investigated if the presence of further potential helpers could influence the sense of obligation. In particular, they predicted that if the subject is asked to imagine to be able to help a victim, knowing that he is relatively closer to the victim than other potential helpers will raise his sense of obligation, while the same would not happen if the subject knows that other potential helpers are in a better position to help. In short, relative distance of potential interveners can influence the sense of obligation. Consider a focal agent P which represents the point of view of the participant, a victim V and a further potential helper A. The comparison here is between four main cases:

1. *Near absolute:* P is the only one who can help V and is near to him.
2. *Far absolute:* P is the only one who can help V and is a bit far from him.
3. *Near relative different:* P and A can both help V, but P is closer to V than A.
4. *Far relative different:* P and A can both help V, but A is closer to V than P.

To be sure that the simple presence of another helper didn't influence the participant motivation, the authors also controlled for *relative same* conditions, where the distance of P and A to V was the same. The hypothesis is that the near/far distinction in absolute cases would not lead to a difference in the obligation to help participants ascribed to P. The sense of obligation would instead be modulated by relative distance, being highest in the *near relative different* case and lowest in the *far relative different* case. Each participant in the experiment had to read a little story about a focal agent who could help a victim from a thief, with or without the presence of a further potential helper who was closer to the victim than the protagonist or further away than him. As predicted, almost the same level of obligation was ascribed to the protagonist by participants who read the absolute versions of the story, while participants who read versions where the protagonist was closer to the victim than the other agent ascribed a greater obligation than those who read version where the protagonist was further away than the victim.

The experiments by Nagel and Waldman do not rule out the possibility that distance might affect our moral judgments. The point is that distance is often accompanied by many other factors that have related effects on moral intuitions. The authors (p. 249) consider these factors to constitute a "family resemblance structure". Expanding on their intuition, I would claim that the factors of a moral problem (whether a real-life problem or a mental experiment) to which human intuition react are quite fuzzy and difficult to pinpoint precisely. They will depend on the particular point of view of the individual person and the

education she has received. Greene (2008, p. 64) believes that consequentialism, and the related reasoning processes, is "systematic and aggregative" and that it can take "everything into account". On the contrary, intuition is set off by the recognition of precise and circumscribed factors. While there is a difference in the way in which conscious cognition and intuition operate (as evidenced by the potential limits to the automation of cognition we looked at in the previous chapter), I don't think that intuition is set off by specific factors, and that it too can be "systematic and aggregative", just in a different way.

### 5.4. THE INDIRECT ROUTE AND DEONTOLOGY

Greene (2014, p. 713) believes that his indirect route argument is an expansion of his direct route argument, and that it constitutes a more general explanation of the reason why some moral judgments are led astray in some contexts. However, the idea that the indirect route is just an expansion of the direct route argument is debatable (Königs, 2018). The first thing to notice is how these arguments have different conclusions. The indirect route argument allows for the possibility that deontological judgments might be more apt to tackle moral problems in certain contexts (when the problem at hand is a familiar moral problem), while such conciliatory conclusion is not present in the direct route argument. This difference in the conclusion of the arguments conceals a deeper difference in the way in which these arguments operate. Stating that moral intuition responds to morally irrelevant factors means trying to investigate which kinds of moral goals one ought to pursue. If a judgment is not the product of a psychological process responding to morally irrelevant factors, then that judgment leads to correct moral conclusions. If instead a judgment is the product of processes responding to morally irrelevant factors, then the judgment deviates from what is rationally valuable. On the other hand, in the indirect route argument, which moral goals are to be pursued seems to be already fixed. The problem with employing deontological judgments in unfamiliar moral problems is that the intuitive processes grounding these judgments are not working in the context they were selected for, that is they are not working on a problem for which we have evolutionary-cultural-personal trial-and-error experience, leading to tribalism and strife among human groups as they are divided by religion, politics, culture, and alike. With his indirect route argument, Greene is implying that the employment of consequentialist-deliberative judgments in unfamiliar moral problems is the cure to tribalism and a way to greater social cohesion, which is instead prevented when we use intuition on those problems. The direct route starts with an investigation on psychological processes (which processes respond to morally irrelevant factors and which do not) to yield a conclusion on moral goals, concluding that the consequentialist goal to minimize suffering is the only one which is rationally worthy to pursue, because it is supported by psychological processes which do not respond to morally

irrelevant factors. On the contrary, the indirect route starts by assuming that given moral goals (like avoiding tribalism and promoting social cohesion) are by themselves good and desirable, and it investigates which kind of moral norms (deontological or consequentialist), and which kind of psychological processes (intuition or deliberation) are best to reach those goals, given the peculiar nature of the moral problem at hand.

One major problem that the indirect route argument faces is that, as it implies that certain moral goals are by themselves desirable, it risks being a circular argument. We might agree, generally, that tribalism is to be avoided and that social cohesion is a good moral end. However, the indirect route argument gives us no reason to prefer a specific solution to a given moral problem, or to prefer one specific moral philosophy to another. The main question within the indirect route argument is not which one between deontology or consequentialism is best at facing a given moral problem, but which between deontology and consequentialism best promotes social cohesion. According to the argument, deontological judgments tackle familiar moral problems with speed and efficiency, but they lead to tribalism and strife among human groups when facing unfamiliar moral problems. The problem with the argument is that the differences between deontology and consequentialism are themselves one of the causes of moral tribalism among human groups. People don't agree on which problems are better solved by consequentialist reasoning and which are better solved by applying deontological considerations, and nothing in the indirect route argument tells us which one is better. The argument just says that when facing certain moral problems deontological judgments (insofar as they are intuitive) are more conducive to strife than the more cognitive consequentialist judgments. But in reality, tribalism is also caused by the fact that not everyone agrees on which between deontology and consequentialism ought to be applied to any given moral problem, and no reason is provided for why consequentialism should be the preferred option.

Consider the following argument: deontological judgments are based on intuitions, and intuitions are set to provide bad kinds of responses when they are applied to an environment (that of unfamiliar moral problems) they were not selected to deal with. One problem with this argument is that, as we saw in the previous sections of this chapter, I suspect that the process of formation of moral intuitions is informed to a much greater degree by rational processes than Greene believes and thus can pick up on a larger variety of factors and problems than the dual-process theory posits. As I have tried to argue, intuition can take into account a variety of factors within traditional philosophical dilemmas like those presented in the trolley problem. The reason why our intuitive responses to such problems present inconsistencies and seem to be set off by morally irrelevant factors, like the presence of personal harm, is due to the highly artificial nature of the problems, where

all potentially relevant factors except personal violence are deemed non pertinent to the mental experiment. I should also note that if the indirect route argument would just concern philosophical mental experiments, the results obtained would be quite trivial: that intuition is set to fail against artificial problems which are designed to be *counterintuitive* sounds almost like a tautology. Moral intuitions were designed to be effective ways to attune us to our social environment, and in the outmost majority of cases we do not face social problems like the footbridge dilemma. But if moral intuitions were found to be good guides in most of our daily lives, as I argue they are, the conclusion that one can design artificial problems which trick our intuitions seems nothing more than a platitude. A further problem stands in the idea that as intuition has not been selected to deal with unfamiliar moral problems, its responses are doomed to be "bad" ones. The problem with this argument is that, as we saw previously, is we define "bad" in terms of what increases tribalism and reduces social cohesion the entire argument starts to be question-begging: one reason why there is tribalism is the unresolved strife between the principles of deontology and those of consequentialism, so, until we have not an argument to prefer one to the other, we cannot use one or the other as a solution to tribalism.

### 5.5. THE INDIRECT ROUTE AND CONSEQUENTIALISM

A related major problem of the indirect route argument is that it seems not to work as a positive case for consequentialism. According to the indirect route, intuitions are set to fail against Us vs. Them problems. These are real-world problems people confront in their daily lives. The feature of these problems making them unfamiliar is the fact that they are recent problems, for which we do not have any common intuition deriving from recurring social problems in our phylogenesis. Everyone will form different intuitions about this kind of problem, relating to their culture or individual experience. Thus, employing reasoning in these circumstances would allow us to bridge the gap between individual contrasting intuitions. My argument here is twofold. For one, as moral intuitions are educated intuitions, they are shaped by individual and collective thinking and thus can constitute an updated mechanism that faces modern problems. Our intuitive gut feelings when confronted with Us vs. Them problems pick up on what makes these problems relevant in the modern world. They are not just a relic. While I agree with Greene that many of these intuitions (but not all) might constitute deontological judgments, being intuitive and being deontological does not prevent them to be up-to-date and complex. Second, I believe that deliberation cannot even operate if not from an evaluative standpoint, which is necessarily intuitive in this kind of case. Take problems like abortion, immigration, euthanasia, the death penalty, taxation, gay marriage. These are all recent problems in which people are divided by politics, religion, social class and so on. Suppose one would not trust one's

intuitions on these problems and try to apply a strict consequentialist approach to them, calculating which kind of choice leads to the best outcome for the greatest number of people. There are two problems with this approach: first, it is often extremely difficult to calculate different effects of different choices on this kind of problem. How many immigrant people would die or be hurt or suffer if we apply a strict anti-immigration policy? How many of our citizens would be hurt (in terms, for example, of increased crime rate) if we allow too many immigrants in our country? There are no clear answers to this kind of question, because they tend to involve so many people in so many disparate situations that even the best consequentialist calculator would struggle to have access to all relevant information and keep track of every possible consequence.

But I would argue that employing a purely rationalist-consequentialist approach to this kind of problem is not only extremely difficult in practice but also theoretically *impossible*. Take abortion, for example. What counts more: the right of the child to be born or the right of the mother to interrupt her pregnancy? This is a purely deontological kind of problem (indeed, it comes spontaneous to use the term "right" in formulating the question), about which people have different kinds of intuitions. But responding to this question first is necessary to calculate the consequences of possible choices. Without having an intuition about the answer to this question, it is impossible to start reasoning at all. Or take taxation: what counts more, the right of people not to be taxed on the fruit of their labor or the right of needy people to be helped by public services that require taxation to be run? Again, without responding to this question, it is impossible to start any calculation about what the greatest good for the greatest amount of people is. The most important moral problems we face are problems where we first need to choose what counts as good (economic freedom or public services, freedom of abortion or right to be born, freedom to move abroad or ethnic identity of a country and so on…) and only then we can calculate, when possible, how to maximize that good. The reason why "Us vs. Them" problems are so difficult is that people disagree on the intuitions about what is good in any given situation, intuitions which are the necessary precondition for any kind of consequentialist calculation.

How is the basic deontological assessment reached? In part it might be fancy philosophizing: a theory of duty and rights which tries to coherently and logically explain why the freedom of the mother counts more of the life of the fetus or vice versa, or why the right of the individual not to be taxed counts more of the right of the homeless person to receive some help or vice versa. But I suspect that even intuition in these cases presents a certain degree of complexity. It reacts not to a single factor, but to an overall story. A narrative, for example, on how women should be freed and emancipated or on how God has given a soul to every human being since inception. An intuition I can have on a specific

case (for example, a woman being denied the right to abort) will not present this level of nuance. I might feel the intuition as a gut feeling of anger. But that intuition reflects the education of intuitions I have been engaged in since my childhood in my specific social environment. As soon as someone asks me to justify my intuition, I reason. This reasoning is probably not just a confabulation defending my intuition (like Haidt believes) nor a mere rationalization of archaic feelings (like Greene seems to believe) but is a process which can express the subconscious reasons motivating my intuition. And if someone makes me notice an inconsistency in my reasoning, I might modify or reject said intuition. Over time, this would lead to intuitions which, while fast and intuitive, can pick up on modern features of a problem. These intuitions will differ from the intuitions reached by other people because their process of education of intuitions will be different from mine. It is improbable that reasoning will help us bridge the gap between different people, because said reasoning will start from another evaluative standpoint. This is what makes this kind of problem so hard to tackle.

# 6. THE ATTUNEMENT MODEL OF MORAL DECISION-MAKING

> "From such crooked wood as that which man is made of, nothing straight can be fashioned."
>
> - *Immanuel Kant*

As we saw in Chapter 1, the Central Tension Principle holds that the distinctions between deontology and consequentialism, cognition and emotion, and intuition and deliberation can be mapped on one another. One can interpret the work we have pursued so far as an attempt to disentangle such union of conceptual distinctions. In Chapter 3, we have seen how the divide between emotion and cognition cannot be meshed with that between intuition and deliberation, and how, more generally, it is a bad framework by which to interpret moral decision-making. In Chapter 4, we have seen how the distinction itself between intuition and deliberation is less strict as previously thought, as education of intuition theories posit that intuitions come to reflect the output of metacognitive deliberation on one previous intuitions. In Chapter 5, we have seen how this distinction cannot even be founded upon the idea that intuition responds to morally irrelevant factors. Despite these conclusions, I believe that there is a difference in the role of intuition and deliberation in moral judgment, just one that is not easily captured by distinctions such as deontology vs. consequentialism, cognition vs. emotion, morally irrelevant vs. morally relevant. In this chapter, I will try to make a positive proposal for a meaningful distinction between intuition and deliberation in moral judgment. In doing so, I will try to lay down the groundwork for a new dual-process theory of moral judgment. Compared to the others, this chapter contains more original ideas, and I fear that it will sound less organic than the rest. However, I hope that by the end of it, the reader will get the gist of the original proposal that motivates my work so far.

## 6.1. THE EDUCATION OF INTUITIONS AS SKILL ACQUISITION

The first part of my original proposal concerns a potential improvement of education of intuitions theories. Sauer (2012a) frames educated intuitions as habits, but in animal cognition research habits are seen as quite recalcitrant to rational influence. As we saw in Chapter 2, a rat that has acquired the habit of pressing a lever in exchange for food will often continue to press the lever even when he shows no more interest in the food (Dickinson et al., 1995). While the notion of habit can be redefined, appealing to the usual notion can be problematic because it would picture a view of educated intuitions as a recalcitrant emotional-intuitive process, rather than a process which is both intuitive and rational. A further potential limitation is constituted by the fact that, to my knowledge, theories of the

education of intuitions tend to leave the extent of direct influence of genuine, explicit, and deliberate reasoning over moral decision-making unspecified. Studies by Haidt and colleagues have shown that the instances where explicit deliberation of a person has a direct effect over their moral convictions are quite rare, but even social intuitionism allows for the possibility that, in rare circumstances, overt deliberation can lead to a suppression or modification of intuitions. Analogously, it would be quite weird to find out that *all* our moral judgments are educated intuitions, leaving no room for explicit reasoning to influence our decision-making. After all, it is reasonable to suppose that some form of moral reasoning cannot be habitualized, just like complex forms of reasoning in other fields cannot be automatized. A theory of the education of moral intuitions should provide us with some method to assess the limits of the automatization of reasoning. In this section, I will argue that conceiving moral agency as a *skill* provides us with a model of the education of intuitions that allows us to securely classify educated intuitions as both rational and intuitive, and that gives us a way to assess the limits of automation.

Skill acquisition is a sophisticated form of self-regulation of behavior we engage in to achieve a desired goal in a domain of high complexity. In acquiring a skill, one internalizes standards about what counts as good performance, which guides one's learning efforts. Achieving expertise in a domain of high complexity requires the completion of many hierarchically organized subgoals. To become better at playing chess one must strive to become better at openings, which in turn can be broken up into many subgoals, which I will refer to as *component abilities*. The progressive mastering of component abilities requires a lot of practice in the form not only of repetition but, most importantly, of *deliberate practice:* the novice needs to continually strive to do things she currently finds difficult. Deliberate practice is also needed to retain expertise, as even competent individuals tend to lose part of their expert performance if they don't train with sufficient regularity. During the acquisition of a skill, effortful tasks become effortless, and attention is freed to handle more complicated tasks (Stichter, 2018, chap. 1). Stichter (2018, p. 27) writes that: "Practice allows us to make progress on tackling ever more difficult tasks by tapping into automaticity."

According to some authors (like Dreyfus and Dreyfus, 1991) skill acquisition implies a global reduction of cognitive control over action, with explicit cognition making no contribution to advanced skilled performance. Christensen, Sutton, and McIlwain (2016) instead argue that during skill acquisition the automation of performance is only partial: automation allows the skilled subject to focus his attention on further aspects of the action carrying them out with greater flexibility and control, having automated lower-level components of the action. The authors argue that *partial automation* better explains a set of

features of skill experience. They focus on these nine features (Christensen, Sutton and McIlwain, pp. 45-46):

i. Attention to performance can be reduced once skill has been acquired (*reduced attention*) – For example, an expert driver can drive without thinking about changing gear when the necessity arises.

ii. A developed skill can be performed in conjunction with other tasks with little detriment (*multi-task tolerance*) – An expert driver can hold a conversation with a passenger while she drives without consequences for the quality of her driving.

iii. Attention to performance can actually disrupt the execution of an acquired skill (*disruptive attention*) – Focusing on the movements required to change gear can lead to a worse execution than doing so automatically.

iv. During the execution of skilled action, the sense of cognitive effort is low (*reduced cognitive effort*)

v. Memory for the performance of skilled action is reduced (*reduced memory*)

vi. Skilled individuals show enhanced attention to strategic features of a task (*strategic focus*) – While the experienced driver has not to focus on basic implementation like changing gears, she shows enhanced attention to things like the route to follow to get to destination.

vii. If the skilled individual does not pay attention, he might perform the wrong action (*action slip*) – While an expert individual is rarely mistaken on the actions she has automated, if she does not pay attention she can still fail at a strategic level. For example, an absent-minded expert driver could turn as if to drive to work when the goal is to go shopping.

viii. Although attention is low in familiar circumstances, it can be very high when faced with demanding circumstances (*increased attention in response to challenge*) – While an experienced driver as no need to focus when driving in a familiar environment, when the situation becomes difficulty the attention of the expert might be very high.

ix. Increased attention when faced with difficult situations can be accompanied by increased sense of cognitive effort (*increased cognitive effort in response to challenge*).

Christensen, Sutton, and McIlwain note how these features are apparently contradictory: attention and cognitive effort are generally reduced but also sometimes increased in skilled action; attention to execution is disruptive, but lack of attention can also lead to action slips. They argue that the reason for these contradictions is that, in skill

acquisition, different aspects of performance are automated. Such a hybrid account has a better chance to account for features like strategic focus, action slip, and increased attention and cognitive effort in response to challenge - which theories predicting full automation of skilled performance seem unable to explain.

The basic idea of Christensen, Sutton, and McIlwain is that automation of behavior tends to be higher for implementation control while strategic and situational control are influenced by cognition. They write (p. 49, italics in the original):

> "*Higher strategic control* involves overall control of the primary skill in relation to its goals. In the case of driving this includes navigation to the destination. *Situation control* involves the control of action in relation to the immediate situation. In the case of driving this involves proximal control of the car in relation to features of the situation, including maneuvers like accelerating to traffic speed, maintaining lane position, maintaining a safe distance to other cars, changing lanes, and so on. *Implementation control* involves performing actions that achieve situation control, which in the case of driving includes steering, accelerating, braking, changing gears, and so on."

The reason why implementation control can be more easily automated is that it involves relatively stable relations, while the relation of action to context is usually more complex and variable. Cognitive control manages the variable features of the situation. It is important to note that we will have two kinds of situations: there are cases where the relation between action and context is highly predictable (for example, driving on the usual route to work when traffic and weather conditions are good) and there are more difficult and unusual situations (for example, driving on at night on a busy highway), I will call these respectively *easy conditions* and *hard conditions*. In both cases, skill acquisition implies the automation of the implementation control, which is constituted by a series of low-level component abilities (for example, changing gear, steering, accelerating and braking…). In easy conditions, automation can spread also to situation control and partially to higher strategic control; while in hard conditions, cognitive control is needed for good performance in the situational and strategic level. So, the automation of skill acquisition will apply to the largest extent to implementation control, and to a lesser degree to situation and strategic control, where such degree will depend on the kind of conditions one is facing. Once implementation is taken care of via automatic processes, attention can be freed up for situation and strategic control in those circumstances where they are only partially automated. The expert will thus be less reliant on attention in general when facing easy conditions, but he will show an

increased attention in hard conditions, with such attention being employed for the strategic control that is needed in those situations.

Here is an explanation of the features of skill experience from the perspective of the hybrid account:

i. *Reduced attention*. Skill acquisition implies a reduction of the attention devoted to low-level implantation details. For example, becoming an expert driver implies the automation of actions such as changing gear. In chess playing, it is often said that the master player does not "see" wrong moves just like the novice does not "see" illegal moves. The basic tactical level is automated, so that the master does not even consider putting the queen on a square that is controlled by the enemy bishop, when there are no clear positional advantages in doing so. But expert performance also implies an overall reduction in attention, even at a strategic level, in easy conditions, where the relation between action and the environment is highly predictable. So, an experienced driver can pay relatively little attention when driving on a familiar route. Similarly, master chess players have automated their openings and can deal with the first moves without paying much attention.

ii. *Multi-task tolerance.* Multi-task tolerance takes two forms. In implementation control, many component abilities present multi-task tolerance, because it is often necessary to perform them together to act efficiently. So, steering, changing gear, braking and so on are component abilities at the level of implementation that the expert driver can perform in rapid sequences when the necessity arises. But the overall skill too can present multi-task tolerance. So, an expert driver can hold a conversation while driving. The multi-task tolerance of a skill will depend on the situation: in easy conditions, the strategic level is also automated to a significant degree, and attention can be freed to further tasks; while in hard conditions, attention must be devoted to the strategic level, and the skill is not multi-task tolerant. For example, an experienced driver can hold a conversation while on a familiar route, but he cannot do the same while performing a difficult maneuver.

iii. *Disruptive attention*. Attention becomes disruptive when it is misdirected from the strategic level to the details of implementation, which have been automated.

iv. *Reduced cognitive effort.* Skill acquisition automates implementation, and it also produces cognitive structures that suit well the demands of the task, reducing cognitive effort. The reduced sense of effort will thus involve both the implementation level and the overall skill in easy conditions.

v. *Reduced memory*. Reduced memory concerns both a reduction of the memory for the details of implementation, and memories related to the overall skill in easy conditions. Memory, however, is not just dependent on cognitive control and attention, as there can be instances where some basic form of cognitive control is present and yet no memory is formed. Memory encoding is affected not just by attention but by the relevance of information for future control (Christensen, Sutton, and McIlwain, 2016, p. 51). In general, there will be more to learn in hard conditions than in easy ones, so situational information in hard conditions will be more likely to be relevant in the future. The authors also hypothesize that experts also possess more fine-grained mechanisms to choose which information is relevant for future control and thus is worth encoding in memory.

vi. *Strategic focus*. By automating implementation, skill acquisition frees attention that can be directed to the management of the strategic level of performance (especially in hard conditions) or to skill-unrelated tasks (especially in easy conditions). For example, by automating the openings and all the other concurrent abilities needed for a basic chess play, the chess master can spare her attention on the strategy needed to checkmate the opponent from a difficult position (this would be the case where attention is freed from the implementation level to be dedicated to the strategic level within the same skill). Or consider the experienced driver that, having automated changing gears and alike, can spare his attention to hold a conversation while driving (this is the case where attention is freed from implementation to be dedicated to a concurrent task external to the skill). Strategic focus is thus compatible with reduced attention because automation of the implementation level allows for attention to be redirected to the situational and strategic one. The redirection of attention to the strategic level means that, for example, an expert driver can be acutely aware of nearby cars during a passing maneuver while at the same time not being attentive to changing gear.

vii. *Action slip*. Action slips happen when the implementation level has been automated but there is weak cognitive control at the strategic level. Implementation control remains mostly untouched, but the relation of the action to its goals becomes misaligned.

viii. *Increased attention in response to challenge.* In hard conditions, the relation between action and the environment tends to be especially complex and

variable and thus the expert needs to use his attention to manage the situation at a strategic level, while the implementation level remains automatic.

ix. *Increased cognitive effort in response to challenge.* Maintaining awareness on the strategic level in hard conditions is experienced as effortful. At the same time the implementation level, which has been automated, is experienced as effortless.

If moral agency can be truly characterized as a skill and if the hybrid account of skilled performance is true, the features of skill acquisition will allow us to solve the problem of the extent of overt reasoning in moral agency. The crucial point is that the automation of performance in skill acquisition is only partial. Automation concerns primarily the lower components of action (component abilities) which are needed for implementation. Once those are automated, attention and cognitive control are freed to focus on the strategic aspects of the skill or on performing concurrent tasks. The strategic level itself might be automated, but the degree by which this automation is successful will depend on whether the expert is facing an easy or hard situation.

The feature of skill acquisition which instead accounts for the rationality of actions even when the subject cannot fully retrieve them is that experts often have difficulty in verbalizing their own skill. Ask a chess master why she made a given brilliant move and sometimes she might provide little explanation besides that she felt that was the right move. The inability of the expert to fully express herself shows how much skilled performance is contained in practice. This feature of skill acquisition (on which Stitcher focuses in section 6 of the first chapter) allows us to drop the requirement that, in order for an action to be rational, the subconscious reasons that were present at the time of actuation must be retrievable from the subject with his post-hoc reasoning. We wouldn't exclude a brilliant chess move from the rank of rational action just because the player who made it cannot fully verbalize why she did. So, if moral agency is a skill, we wouldn't exclude many moral actions from being rational just because the person who made them couldn't fully verbalize why he did.

In this section, we have looked at how the skill model of moral decision-making would overcome some limitations of traditional theories of the education of intuitions, allowing us to: (1) characterize moral agency as both rational and intuitive; (2) account for the role of explicit non-post-hoc reasoning in moral decision-making; (3) account for the rationality of moral actions even when subjects cannot fully verbalize the reasons for them.

### 6.2. THE BANAL DUAL-PROCESS MODEL OF MORAL DECISION-MAKING

The second component of my positive proposal is a critique of the idea that consequentialist judgments are not related to greater deliberative capacity. Kahane (2014)

notices how, in the personal and impersonal moral dilemmas usually employed by Greene, the consequentialist judgments (like pushing the man in the footbridge dilemma) are usually also highly counterintuitive judgments, while deontological judgments are strongly intuitive, and wonders whether the behavioral and neurological differences observed in subject performing either one or the other judgment might reflect a difference between the processing of intuitive judgments and counterintuitive ones rather than a difference between deontology and consequentialism. Here the use of the term "intuitive" applies to the judgment itself, and not to the psychological process leading to the judgment. I think that employing the term "intuitive" for both the process leading to a judgment and a property of the judgment itself might be confusing. The reason is that, while in most cases reaching a counterintuitive judgment might involve deliberation, in most actual instances of reasoning deliberation might just confirm the intuitive judgment. While Greene posits that intuition produces responses which often differ from deliberation, both deliberation and intuition often agree that an intuitive moral judgment is correct. To make an example: I might find it intuitively correct that killing innocent people is wrong, and my reasoning agrees with this intuitive moral judgment. For this reason, in the following I will talk of *common* moral judgments and *uncommon* ones. Roughly, one moral judgment is common if most people agree that it is right, while uncommon moral judgments are those that at first most people consider to be wrong. Obviously, there will be much cultural variation between those judgments that are taken as common within each human society. In the last section of this chapter, it will become clear why I think that this very rough distinction between common moral judgments and uncommon ones cannot be mapped onto the distinction between familiar moral judgments and unfamiliar ones as made by Greene. For the moment, let us return to Kahane's reasoning: the behavioral and neurological differences registered by Greene might not be tracking the distinction between deontology and consequentialism but that between common and uncommon moral judgments (intuitive and counterintuitive judgments, in Kahane's jargon).

The crucial observation here is that not all uncommon judgements are consequentialist, and not all common judgments are deontological: the consequentialist judgment that one should lie in order to save a life strikes as common for most people, while the contrary deontological judgment is really uncommon, to the point that even the fancy philosophizing of a great thinker like Kant to defend it sounds unconvincing. More generally, as we have seen in Chapter 1, consequentialist judgments with a high kill-to-save ratio tend to be highly common and even performed automatically. If one takes the Central Tension Principle as it is, one will predict that consequentialist judgments should involve deliberate processing, whether they are common or not, and that the neural and behavioral

correlates of uncommon deontological judgments (like the Kantian judgment on the impermissibility of lying even to save lives) would significantly differ from the correlates of uncommon consequentialist judgments (like the judgment that it is right to push the person of the footbridge). Contrary to this prediction, Kahane (2014) predicts that uncommon judgements, regardless of whether they are consequentialist or deontological, should present similar correlates, that would differ from the correlates of common judgments. The hypothesis is thus that there is a significant difference between judgments according to their *commonness* ("intuitiveness" in Kahane's jargon) but there is no significant difference between judgments according to their *content* (whether they are deontological or consequentialist).

To test this hypothesis, Kahane and colleagues (2012) investigated the neural and behavioral correlates of dilemmas where the deontological answer is common vs. dilemmas where the consequentialist answer is common. They found no difference in difficulty rating and response time between consequentialist and deontological judgments, but a significant difference in the difficulty rating between common and uncommon judgments, with the latter being perceived as more difficult. Furthermore, consequentialist judgments such as finding it right to push the person in the footbridge dilemma and deontological judgments such as finding it wrong to lie even in case the lie could save lives recruited similar brain areas, while the corresponding common judgments were different under this profile. These results lead to two conclusions: first, the neural and behavioral correlates of choices like pushing the person in the footbridge dilemma might not reflect any consequentialist calculation or the application of consequentialist principles, but rather be a consequence of the fact that these choices are uncommon; second, they hint towards a "banal" dual-process model of moral judgment where common judgments are associated with a faster, more automatic processing (intuition), while uncommon judgments, regardless of whether they are deontological or consequentialist, are supported by slower, more deliberate processing (Kahane, 2014).

THE "BANAL" DUAL-PROCESS MODEL

Common moral judgments are generally supported by fast and automatic processing (intuition), while uncommon moral judgments are preferentially associated with conscious reasoning and allied processes of cognitive control (deliberation).

### 6.3. WHAT PATHOLOGY TELLS US ABOUT MORAL JUDGMENT

As we saw in Chapter 1, the Central Tension Principle is supported by a series of evidence coming from brain imaging and cognitive load. Greene also reports a series of data

from pathology purportedly supporting dual-process interpretation and in particular the Central Tension Principle (Greene, 2014, pp. 701-704). This literature is quite extensive. Here, I will focus on what is arguably one of the main pieces of this evidence: the results of the damage to the ventromedial prefrontal cortex (vmPFC). The vmPFC is usually associated with the production of emotion. Koenigs and colleagues (2007, p. 908) write that patients with lesions to the vmPFC usually exhibit:

> "[…] generally diminished emotional responsivity and markedly reduced social emotions (for example, compassion, shame and guilt) that are closely associated with moral values, and also exhibit poorly regulated anger and frustration tolerance in certain circumstances. Despite these patent defects both in emotional response and emotion regulation, the capacities for general intelligence, logical reasoning, and declarative knowledge of social and moral norms are preserved."

Damage to the vmPFC causes a general impairment in motivational states, which ruins practical real-life decision-making. Patients who acquire vmPFC damage as children also tend to develop sociopathic traits. This fact has been used to argue that the proper development of social emotions is essential to acquire a moral sense (Anderson et al., 1999). If emotions, mediated by the vmPFC, exert a critical influence on moral judgment, individuals with vmPFC lesions should exhibit an abnormally high rate of consequentialist judgments on emotionally salient moral scenarios (like the footbridge dilemma), but a normal pattern of judgment on less emotional moral scenarios (like the trolley dilemma) (Koenigs et al., 2007, p. 908). Koenigs and colleagues found that vmPFC patients did choose to take action on consequentialist grounds more often than the control group, but only when the conflict between consequentialist considerations and the emotional (deontological) response was high (like in the footbridge dilemma). When faced with low-conflict moral scenarios (recall the infanticide dilemma), the vmPFC patients' responses were consistent with those of the control group. This is coherent with an interpretation according to which these patients can rely on explicit knowledge of social and moral norms which prohibit doing harm to others but have impaired emotional reactions to emotionally salient actions and events. This suggests that damage to the vmPFC does not cause a general impairment of capacity for moral judgment, nor leads to always prefer consequentialist considerations, but is critical "only for moral judgments in which social emotions play a pivotal role in resolving moral conflict" (Koenings et al., 2007, p. 910). Greene takes these results on vmPFC patients as evidence for the idea that emotional reactions are related to deontological judgments.

Greene's interpretation of the function of the vmPFC is challenged by a behavioral study by Koenigs and Tranel (2007). In the Ultimatum Game, two players are given the opportunity to split a sum of money. One player (the proposer) can share whichever proportion of money he wants with the second player (the responder). If the responder accepts the offer, the players split the money according to the offer made by the proposer. If the responder refuses the offer, both players receive nothing. According to game-theoretical norms of rationality, the responder should accept whichever offer, no matter how skewed in favor of the proposer, because the alternative is receiving nothing. However, it is a robust finding in empirical psychology that "unfair" offers (in which the proposer retains a significantly higher proportion of money) tend to be rejected by the responder (Guth, Schmittenberger & Schwarze, 1982); a rejection which is associated with emotions like anger (Pillutla & Murnighan, 1996). Koenigs and Tranel have found that vmPFC patients present a higher tendency to reject unfair offers, which can be interpreted as the patients providing an overly "emotional" response. This pattern of response is difficult to explain from a perspective like that of Greene, which considers these patients to be more cognitive given the damage to a brain area related to emotional processing.

Selective damage to the vmPFC causes both an increase in consequentialist judgments when the subject is confronted with personal moral dilemmas, and an increase of rejection of unfair offers in the Ultimatum Game, which is usually due to an increased emotional response. One proposed explanation for these results is that the vmPFC, rather than being linked to the production of emotions, is involved in the creation of prosocial *moral sentiments*, with damage to the area causing an increase in self-centered and other-aversive emotional experience (Moll & de Oliveira-Souza, 2007). This would explain the "colder" response to personal moral dilemmas and the more egoistic choices in the Ultimatum Game. Crucially, these prosocial sentiments involve both affective and cognitive aspects and can be thought of as the refined kind of intuitions which figure in the education of intuitions theories. The peculiar pattern of response we obtain from patients with lesions to the vmPFC in the sacrificial dilemmas and the Ultimatum Game, a pattern which shows a heightening of egoistic behavior, illuminates the goal of prosocial sentiments: to promote the *attunement* of individuals to the subtleties of social life, and their impairment heightens egoistic behavior (de Oliveira-Souza, Moll & Grafman, 2011). The model of prosocial sentiments (which finds an ancestor in Adam Smith's, 1759/1976, theory of moral sentiments) allows thus to trace a distinction between self-centered and other-centered behavior, but not a distinction between purely cognitive and purely emotional judgments. While standard approach theories distinguish emotion and cognition at the level of their neurocognitive functioning and their phylogenetic origin, the attunement approach considers the fine-

tuning of the individual to the needs of social life to be characterized at multiple levels by an inextricable union of emotional and cognitive aspects. It predicts that damage to specific components of the network producing such attunement will not cause a selective damage to either reason or emotion as cohesive and distinct faculties, but rather an impairment of the ability of the patient to regulate his emotional and cognitive processes in accordance with societal expectations.

### 6.4. COMMONNESS, SELF-/OTHER-CENTEREDNESS, AND CONSEQUENTIALISM

So far, we have seen a distinction between common moral judgments and uncommon moral judgments, and a distinction between self-centered and other-centered behaviors and thoughts. In this section, I want to respond to three interrelated questions. First, assuming that the banal dual-process model of moral judgment is correct, what is shared among common-intuitive moral judgments on the one side and among uncommon moral judgments on the other, if it is not a difference between deontology and consequentialism? Second, what is the relation between the two distinctions discussed so far (common/uncommon, self-centered/other-centered)? Third, following Kahane and colleagues (2018), I want to investigate a potential mystery related to consequentialism.

Consider the first question: what is shared by common moral judgments vs. uncommon moral judgments? In most healthy people, deliberate processing is needed to reach the uncommon consequentialist conclusion that it is right to push a person on the tracks in the footbridge dilemma, and deliberate processing is also needed to reach the uncommon deontological conclusion that it is wrong to lie even when this would result in the saving of a life. That is, in most healthy people, common moral judgments are intuitive, while uncommon moral judgments are deliberate. As I have argued in the previous chapters, this difference between judgments that are universally agreed upon and judgments that people have difficulty to accept even after the arguments by philosophers cannot be mapped upon a difference between schools of ethical thought, nor is it due to a differences in psychological underpinning such as emotion and "cold" cognition, nor can it be attributed to the fact that common judgments react to morally irrelevant factors. So why does the idea of pushing the person of the bridge strike most of us as morally repugnant, and the idea that it is right to lie to save lives as morally correct? As seen in Chapter 5, it is hard to pinpoint exactly what features of a problem intuition is reacting to, but, as Kahane and colleagues' study shows, there is a significant difference between common and uncommon judgments. The hypothesis I want to put forward here is that, at least under many circumstances, the difference follows the line of self-centered vs. other-centered behaviors and thoughts. This does not mean that what common judgments share is the fact that they are other-centered while all uncommon judgments are self-centered. Rather, the

correct attunement of the individual to the needs of social life requires a balance between self-centeredness and other-centeredness. Judgments which present this balance will feel common, while judgments that lean too much towards one of the extremes will feel uncommon.

My hypothesis is that the distinction between other-centeredness and self-centeredness is a classification of moral behavior that captures a true divide present in human moral judgment. I hypothesize that common moral judgments, which are mostly processed intuitively, respect a balance between other-centeredness and self-centeredness, while uncommon moral judgments are those that violate this balance. The basic dimension is thus that of the correct level of attunement to social expectations. Kahane and colleagues (2012) have shown that neural and behavioral correlates of common moral judgments are similar. I would also predict that these judgments are felt to present a right balance between self-centeredness and other-centeredness. What happens in pathology? The balance breaks down. The person becomes, for example, more self-centered (both in his "cold" thoughts and his emotional reactions). While in the healthy subjects uncommon moral judgments require a good deal of deliberation, in pathologically self-centered subjects, like those who received damage to the vmPFC, the uncommon moral judgments become intuitive. As noted by Kahane (2014), the fact that people with damage to the vmPFC produce more consequentialist choices in personal moral dilemmas is not actually evidence that they are *reasoning more* (that is, applying an impartial consequentialist reasoning), rather it is evidence that they are *feeling less* (feeling less constrained by rules and norms to which we normally attribute emotional value). Indeed, for the vmPFC patient the consequentialist judgment is intuitive, and is not associated with any kind of deeper reasoning, application of principles, or calculation about harm (Moretto et al., 2009). Such judgment does not follow from any real consequentialism, but it is done just because it is the more self-centered option. This also explains why vmPFC patient are not only more likely to push the person from the footbridge, but are also more egoistic in the Ultimatum Game, and, generally, more socially aversive in their daily lives. The relevant difference is between being other-centered and being self-centered, and this difference is transversal to the previously discussed distinctions.

So far, one might think that, according to the model proposed here, consequentialist judgment is mostly related to self-centered thoughts and behaviors, that, for example, make vmPFC patients ignore the rules healthy subjects attribute significance to. This is only partially true. I suspect indeed that the rules of consequentialism, taken in their pure value, often strike as the source of uncommon moral judgments, but this does not happen because they always prescribe a behavior that is particularly self-centered. Indeed, the ideal of

classical utilitarianism that we should not show preference for our in-group but extend our sphere of moral concern to all sentient beings is strongly other-centered. I suspect that, both intuitively and on deeper deliberation, most of us don't find such an ideal to be morally wrong, rather most people find it to be *too* morally good, that is *saintly* good. Most people have a natural self-centeredness and a natural preference for their family, friends, compatriots, and alike. When the consequentialist comes and says that we should be impartial towards all human beings, this strikes as an unrealistic maxim to follow. The reason is that it leans too much in the opposite direction than the self-centered nature of vmPFC patients: it is even too other-centered. That the ideals of utilitarianism are perceived as saintly good allows us to get a correct picture of the relation between the distinction between common vs. uncommon moral judgments and the distinction between behaviors and thought that are self-centered vs. those that are other-centered. While the ideals of classical utilitarianism are uncommon for prescribing an extension of our moral compassion which exceeds the boundary of our natural self-centeredness, the consequentialist choices of vmPFC patients are uncommon because they showcase a heightened self-centeredness.

How can the same moral philosophy yield judgments that are perceived as uncommon due to their utter self-centeredness and for their extreme other-centeredness? The reason is that consequentialist judgments in lay people might not be a unitary psychological phenomenon and might reflect two distinct psychological processes (Kahane et al., 2018): on the one side, consequentialism reflects *impartial beneficence*, that is the extent to which a person endorses the impartial promotion of everyone's welfare; on the other side, consequentialism reflects a greater propensity to justify *instrumental harm*, that is the extent to which the subject endorse harm that brings a greater good. The hypothesis put forward by Kahane and colleagues is that the increased consequentialist judgment of clinical populations like vmPFC patients and psychopaths reflects a greater tendency to see instrumental harm as justified and has nothing to do with impartial beneficence. Similarly, empathic concern for others is strongly correlated with impartial beneficence and inversely correlated with the propensity to see instrumental harm as justified. Insofar as it prescribes the application of impartiality in the treatment of others, consequentialism is associated with greater empathy and heightened emotions (we can call this *positive consequentialism*). Insofar as it prescribes that instrumental harm can be performed to achieve greater good, consequentialism is associated with a lack of empathy and stunned emotion (*negative consequentialism*). The extreme forms of either of these tendencies are perceived as somewhat deviant from the average moral perspective: pure forms of impartial beneficence stand out for being saintly good, while an extreme acceptance of instrumental harm strikes as morally repugnant. My own prediction is that most lay people would find these extreme forms of

either side of consequentialism deviant from different reasons, and I suspect that the motive of such a judgment stands in the fact that the extreme form of consequentialism-as-impartial-beneficence is excessively skewed towards other-centeredness, while the extreme form of consequentialism-as-instrumental-harm is excessively skewed towards self-centeredness.

Looking back at the classifications reviewed in previous chapters, a new dual-process model of moral judgment can be now proposed. The crucial distinction within this new model, which I call the *attunement model* of moral decision-making, is that between self-centered and other-centered behaviors and thoughts. Moral judgment at either extreme will be perceived, by most healthy subjects, as uncommon, and will be usually supported by deliberate processes. On the contrary, moral judgments where a balance is found will be perceived as common and will be usually supported by intuitive processes. What it means that intuitive processes support common judgments and that deliberate processes support uncommon judgments is that, while deliberate processes usually agree with intuition that common judgments are right, to deviate from common judgments healthy subjects need deliberation. Common judgments are thus the baseline on which both intuition and deliberation usually converge, but deliberation can also be used to deviate from them. In pathology instead, uncommon judgments become more intuitive: in vmPFC patients for example, judgments which strongly lean towards self-centeredness become more intuitive. In the attunement model, a union is made between three major conceptual divisions: self-centeredness vs. other-centeredness, common judgments vs. uncommon judgments, and intuitive judgments vs. deliberative ones. The relationship between these three conceptual distinctions should appear clear now.

More complex is the relationship between the model so far and three further conceptual distinctions: emotion vs. cognition, deontology vs. consequentialism, and judgements responding to morally irrelevant factors vs. judgments responding to morally relevant ones. As for the first distinction, it should be noted that common moral judgments are both cognitive and emotional at the same time, and the same holds true for uncommon moral judgments. When in pathology uncommon moral judgments become more intuitive, this should not be interpreted as the judgments of patients becoming more emotional or more cognitive, but as the thoughts and feelings of these patients becoming, as in the case of vmPFC patients, more self-centered. As for the deontology-consequentialism distinction, I suspect that consequentialism, in its pure form, always strikes as the source of quite uncommon moral judgments. The reason is that negative consequentialism (greater propensity to justify instrumental harm) is related to strong self-centeredness, while positive consequentialism (impartial beneficence) is related to strong other-centeredness.

This hints at the possibility that deontological principles might, in the judgment of lay people, be less extreme, and be the source of more common moral judgments. This early hypothesis is not to deny that, obviously, some deontological judgments can be uncommon (think of the judgment defended by Kant that one ought not to lie even to save lives). As for the morally relevant/morally irrelevant factors distinction, I think that the entire distinction is misguided. As we saw in Chapter 5, moral intuition responds to all sorts of factors, some of which are more central to morality, while others would seem less "relevant" but still affect our moral judgment. But I think there is no neutral standpoint from which to evaluate what counts as a morally relevant factor. Some philosophers might be repelled by the fact that peoples' moral judgments are affected by factors such as whether the recipient of a good action is a compatriot and a foreigner or similar considerations, but I believe there is no perspective from which we can definitively conclude, by reasoning alone, that such considerations ought not influence our moral judgment.

### 6.5. CONSIDERATIONS ON THE ATTUNEMENT MODEL

Here, I will take some space for further considerations about the attunement model, starting with the following question: if the model is true, would any normative consequences follow from the model, if it is true? Common moral judgments are those judgments which strike us as intuitively right and which, most often, deliberation confirms to be correct. Consequentialism, taken in its pure form, is instead the source of judgments which are often perceived as uncommon, and which require a lot of deliberation to be accepted, in the rare occasions in which they are. The reason is that, in its positive side, consequentialism requires us to overcome the natural self-centeredness of our interests, and this is too big of a leap for most people. There are quite a few people that are willing to perform incredible acts of altruism: from saving the drowning child from a pond to sacrificing themselves on a mine to save comrades. But these acts of incredible altruism are not at all as *impartial* as consequentialism would require, as we far more readily sacrifice ourselves for our family, our loved ones, our friends, our compatriots, and alike, than for strangers or for obscure ideals like humanity as a whole. In most of our moral judgments, we are guided by intuitions which skew our judgments towards our interests and the people we like. Self-centeredness is not altruism, and positive consequentialism, strictly speaking, requires us to renounce the first former rather than the latter. Humans are capable of extraordinary acts of altruism, but most of them tend to be *tribalist* in nature: we would rather sacrifice ourselves for people of our in-group rather than humanity.

The attunement model also proposes that reasoning and deliberation would not at all lead to a correction of this natural tribalist tendency. Indeed, our deliberation often agrees with intuition and leads just to a more complex narrative of why we ought to prefer our

ingroup rather than the outgroup. I suspect that in most cases deliberation can actually make tribalism worse by giving it a more reasoned framework. Deliberation can lead from the natural noticing of a difference in skin color to racist theories, it can lead from simple recognition of differences in belief to a theory of why we ought to hate that particular religious group, and so on. Furthermore, as noted by theories like moral disengagement (Bandura, 1999), deliberation can make things worse by working against our natural dispositions against performing harm on one another. We have the intuition that killing innocent people is wrong, a natural deep repulsion to perform gratuitous harm, but we also have the uncanny ability to reason around this intuition by constructing a narrative that they are the inferior race, that they have as a group taken something from our group and so on. Probably there is no easy way to overcome humanity natural tribalism and neither consequentialism nor deeper reasoning would work as solutions. Indeed, sometimes the best solution is to apply some deontological principles ("I ought to respect the outgroup people, no matter how weird their beliefs seem to me") and to trust our deep-seated intuitions about not harming people. Adopting this perspective, it is no wonder that some of the greatest tragedies of recent times have been performed by people following complex ideological systems which told them to suppress their intuitions about what is right or wrong in the name of radical and supposedly rational reforms of how humans should live. The normative proposal of the attunement model thus much in line with the quotation by Kant opening this chapter: we should accept, at least partially, the fact that our moral judgments are skewed towards ourselves and our ingroup, and that this kind of self-centeredness, while being the source of many problems, is also unavoidable, and somewhat desirable. Tribalism is indeed a problem, but there is a healthy dose of self-centeredness in moral matters.

Another question is what relationship is there between the common-uncommon distinction related to moral judgments and Greene's familiar vs. unfamiliar distinction related to moral problems. As defined by Greene, familiarity derives from personal, cultural, or evolutionary trial-and-error experience. Königs (2018) has a series of intelligent questions for the metaphor of Greene's model under this regard, which are generally related to the idea that the processes of trial-and-error experience might not be as reliable in familiar moral contexts. I would add here that processes of cultural and personal learning are much more complex than Greene seems to envision, and they will require a lot of interaction between deliberative processes and intuitive ones. Which moral judgement will be considered to be common will depend a lot on the culture a subject is in and his personal experience with moral matters, and maybe which problems are familiar or unfamiliar is far less clear cut than Greene seems to believe. My own terminology is thus less theoretically loaded: while

Greene's familiar problems are those for which we have sufficient trial-and-error experience, what moral judgment is considered common will depend on that experience. As for the relationship between culture and evolution in the moral sphere, the Moral Foundations Theory (Graham et al., 2013) has some interesting remarks, but it will bring me much further than my present purposes.

The dependence of what counts as a common moral judgment from the subject's culture and personal experience is important to understand the relationship between the model sketched here and the education of intuitions theories. It is possible that what counts as common does not depend purely on intuition or emotion but is slowly constructed through a process of education of intuitions, and, more generally, through a process of moral education of the subject. We construct some early intuitions, and then we reflect on them given the feedback we receive from experience, and progressively we construct a representation of judgments toward which our intuition and deliberation agree. The feedback we receive on our early intuitions will depend on the culture we are in, which itself develops in the boundaries of what is evolutionarily advantageous. I suspect that the role of culture in the development of the educated intuitions in a specific person is much stronger than the role played by evolution, which is just the external constraint. So, morality wouldn't be largely innate, but rather the result of a culturally informed process of personal development. I also think that there is nothing specifically moral about such a process: we form a lot of early intuitions on a series of topics, then, on the basis of culturally mediated feedback, we apply our general reasoning abilities, which, interacting with our intuitions, converge on the judgments that will appear common to us. I think that morality, in the form of common moral judgments, is bootstrapped through domain-general learning mechanisms which include both intuitive and reasoned aspects. A theory of moral learning criticizing earlier domain-specific theories has been recently offered by Shaun Nichols (2021). Surely, integrating the present considerations with that model could reveal itself as a fruitful way forward.

# BIBLIOGRAPHY

Adolphs R. (2010). What does the amygdala contribute to social cognition?. *Annals of the New York Academy of Sciences*, *1191*(1), 42-61.

Adolphs R., Russell J. A. & Tranel D. (1999). A role for the human amygdala in recognizing emotional arousal from unpleasant stimuli. *Psychological Science*, *10*(2), 167-171.

Adolphs R., Tranel D., Damasio H., & Damasio A. R. (1995). Fear and the human amygdala. *Journal of neuroscience*, *15*(9), 5879-5891.

Anderson M. L. (2007a) Evolution of cognitive function via redeployment of brain areas. *The Neuroscientist 13*:13–21.

Anderson M. L. (2007b) The massive redeployment hypothesis and the functional topography of the brain. *Philosophical Psychology 21*(2):143–74

Anderson M. L. (2008) Circuit sharing and the implementation of intelligent systems. *Connection Science 20*(4):239–51.

Anderson M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(4), 245-266.

Anderson S. W., Bechara A., Damasio H., Tranel D. & Damasio A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nat Neurosci 2*:1032-1037.

Ardila A., Bernal B. & Rosselli M. (2016). Why Broca's area damage does not result in classical Broca's aphasia. *Frontiers in human neuroscience*, *10*, 249.

Bago B. & De Neys W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782.

Bandura, A. (1999). Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review*, *3*(3), 193-209.

Barsalou L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Barton R. A. & Venditti C. (2013). Human frontal lobes are not relatively large. *Proceedings of the National Academy of Sciences*, *110*(22), 9001-9006.

Baxter M. G. & Murray E. A. (2002). The amygdala and reward. *Nature reviews neuroscience*, *3*(7), 563-573.

Belova M. A., Paton J. J., Morrison, S. E. & Salzman C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron*, *55*(6), 970-984.

Belova M. A., Paton J. J. & Salzman C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *Journal of Neuroscience*, *28*(40), 10023-10030.

Berker S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, *37*(4), 293-329.

Białek M. & De Neys W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology, 28,* 631–639.

Białek M. & De Neys W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12,* 148–167.

Binkofski F., Amunts K., Stephan K. M., Posse S., Schormann T., Freund H. J., Zilles K. & Seitz R. J. (2000). Broca's region subserves imagery of motion: a combined cytoarchitectonic and fMRI study. *Human brain mapping*, *11*(4), 273-285.

Blackford J. U., Buckholtz J. W., Avery S. N. & Zald D. H. (2010). A unique role for the human amygdala in novelty detection. *Neuroimage*, *50*(3), 1188-1193.

Bless H., Schwarz N. & Kemmelmeier M. (1996). Mood and stereotyping: Affective states and the use of general knowledge structures. *European review of social psychology*, *7*(1), 63-93.

Bodenhausen G. V., Kramer G. P. & Süsser K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of personality and social psychology*, *66*(4), 621.

Boeckx C. & Benítez-Burraco A. (2014). The shape of the human language-ready brain. *Frontiers in psychology*, *5*, 282.

Boyle M. (2016). Additive theories of rationality: A critique. *European Journal of Philosophy*, *24*(3), 527-555.

Breiter H. C., Etcoff N. L., Whalen P. J., Kennedy W. A., Rauch S. L., Buckner R. L., Strauss M. M., Hyman S. E., Rosen B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, *17*(5), 875-887.

Broks P., Young A. W., Maratos E. J., Coffey P. J, Calder A. J., Isaac C. L., Mayes A. R., Hodges J. R., Montaldi D., Cezayirli E., Roberts N., Hadley D. (1998). Face processing impairments after encephalitis: amygdala damage and recognition of fear. *Neuropsychologia*, *36*(1), 59-70.

Campbell R. & Kumar V. (2012). Moral reasoning on the ground. *Ethics*, *122*(2), 273-312.

Craigie J. (2011). Thinking and feeling: Moral deliberation in a dual-process framework. *Philosophical Psychology*, *24*(1), 53-71.

Christensen W., Sutton J. & McIlwain D. J. (2016). Cognition in skilled action: Meshed control and the varieties of skill experience. *Mind & Language*, *31*(1), 37-66.

Cushman F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273-292.

Damasio H., Grabowski T. J., Tranel D., Hichwa R. D. & Damasio A. R. (1996) A neural basis for lexical retrieval. *Nature 380*:499–505

Damasio A. R. & Tranel D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, *90*(11), 4957-4960.

De Caro M., Marraffa M. & Vaccarezza M. S. (2021). The priority of phronesis: How to rescue virtue theory from its critics. In De Caro M. and Vaccarezza M.S. (eds.), *Practical Wisdom* (pp. 29-51). Routledge.

de Oliveira-Souza R., Moll J. & Grafman J. (2011). Emotion and social cognition: Lessons from contemporary human neuroanatomy. *Emotion Review*, *3*(3), 310-312.

de Sousa R. (1987). *The rationality of emotion*. Cambridge: MIT Press.

Dove G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic Bulletin & Review, 23*, 1109–1121

Dreyfus H. L. & Dreyfus S. E. (1991). Towards a phenomenology of ethical expertise. *Human studies*, 229-250.

Dubois S., Rossion B., Schiltz C., Bodart J. M., Michel C., Bruyer R. & Crommelinck M. (1999). Effect of familiarity on the processing of human faces. *Neuroimage*, *9*(3), 278-289.

Dickinson A., Balleine B., Watt A., Gonzalez F. & Boakes R. A. (1995). Motivational control after extended instrumental training. *Learning & Behavior*, *23*, 197-206.

Evans J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*(1), 255-278.

Evans J. S. B. T. & Stanovich K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223-241.

Frankish K. & Evans J. S. B. T (2009). The duality of the mind: An historical perspective. In Evans J. S. B. T. and Frankish K. (eds.), *In two minds: Dual processes and beyond*. Oxford University Press.

Fine C. & Blair R. J. R. (2000). The cognitive and emotional effects of amygdala damage. *Neurocase*, *6*(6), 435-450.

Fodor J. A. (1983). *The modularity of mind*. MIT press.

Garavan H., Pendergrass J. C., Ross T. J., Stein E. A. & Risinger R. C. (2001). Amygdala response to both positively and negatively valenced stimuli. *Neuroreport*, *12*(12), 2779-2783.

Geschwind N. (1970). The Organization of Language and the Brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior. *Science*, *170*(3961), 940-944.

Ghods-Sharifi S., Onge J. R. S. & Floresco S. B. (2009). Fundamental contribution by the basolateral amygdala to different forms of decision making. *Journal of Neuroscience*, *29*(16), 5251-5259.

González J., Barros-Loscertales A., Pulvermüller F., Meseguer V., Sanjuán A., Belloch V. & Ávila C. (2006). Reading cinnamon activates olfactory brain regions. *Neuroimage*, *32*(2), 906-912.

Graham J., Haidt J., Koleva S., Motyl M., Iyer R., Wojcik S. P. & Ditto P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Academic Press.

Greene J. D. (2008). The secret joke of Kant's soul. *Moral psychology*, *3*, 35-79.

Greene J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology*, *45*(3), 581-584.

Greene J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, *124*(4), 695-726.

Greene J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66-77.

Greene J. D., Cushman F. A., Stewart L. E., Lowenberg K., Nystrom L. E. & Cohen J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364-371.

Greene J. D., Morelli S. A., Lowenberg K., Nystrom L. E. & Cohen J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144-1154.

Greene J. D., Nystrom L. E., Engell A. D., Darley J. M. & Cohen J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400.

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.

Gürçay B. & Baron J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, *23*, 49–80.

Guth W., Schmittenberger R. & Schwarze B. (1982). An experimental analysis of ultimatum bargaining. *J Econ Behav Organ 3*:376.

Haidt J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt J. & Björklund F. (2008). Social intuitionists answer six questions about moral psychology. In Sinnott-Armstrong W. (eds.), *Moral Psychology, Vol 2: The Cognitive Science of Morality: Intuition and Diversity*, pp. 181–217. Cambridge, MA, US: MIT Press.

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

Hamann S. & Mao H. (2002). Positive and negative emotional verbal stimuli elicit activity in the left amygdala. *Neuroreport*, *13*(1), 15-19.

Hanoch Y. (2005). One theory to fit them all: The search hypothesis of emotion revisited. *British Journal for Philosophy of Science, 56*, 133–14

Hatfield T., Han J. S., Conley M., Gallagher M. & Holland P. (1996). Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. *Journal of Neuroscience*, *16*(16), 5256-5265.

Hauser M., Cushman F., Young L., Kang-Xing Jin R. & Mikhail J. (2007). A dissociation between moral judgments and justifications. *Mind & language*, *22*(1), 1-21.

Herry C., Bach D. R., Esposito F., Di Salle F., Perrig W. J., Scheffler K., Lüthi A. & Seifritz E. (2007). Processing of temporal unpredictability in human and animal amygdala. *Journal of Neuroscience*, *27*(22), 5958-5966.

Hume, D. (1740/1978). *A treatise of human nature*. Selby-Bigge L. A. and Nidditch P. H. (eds.), Oxford: Oxford University Press.

James W. (1890/1981). *The Principles of Psychology*. Cambridge, Mass.: Harvard University Press

Kaada B. R. (1951). Somato-motor, autonomic and electrocorticographic responses to electrical stimulation of rhinencephalic and other structures in primates, cat, and dog; a study of responses from the limbic, subcallosal, orbito-insular, piriform and temporal cortex, hippocampus-fornix and amygdala. *Acta Physiologica Scandinavica. Supplementum*, *24*(83), 1-262.

Kahane, G. (2014). Intuitive and counterintuitive morality. In D. Jacobson & J. D'Arms (eds.), *The science of ethics: Moral psychology and human agency*. Oxford: Oxford University Press

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological review*, *125*(2), 131.

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.

Kahneman D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kahneman D. & Fredrick S. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. In Gilovich T., Griffin D. and Kahneman D. (eds.), *Heuristics and biases*, pp. 49-81. New York: Cambridge University Press.

Kamm F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. New York, NY: Oxford University Press.

Kant I. (1785/1959). *Foundation of the metaphysics of morals*. Indianapolis: Bobbs-Merrill.

Kapp B. S., Supple Jr. W. F. & Whalen P. J. (1994). Effects of electrical stimulation of the amygdaloid central nucleus on neocortical arousal in the rabbit. *Behavioral neuroscience*, *108*(1), 81.

Kawashima R., Sugiura M., Kato T., Nakamura A., Hatano K., Ito K., Fukuda H., Kojima S. & Nakamura K. (1999). The human amygdala plays an important role in gaze monitoring: A PET study. *Brain*, 122(4), 779-783.

Kiefer M. & Pulvermüller F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825.

Koenigs M. & Tranel D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *Journal of Neuroscience*, *27*(4), 951-956.

Koenigs M., Young L., Adolphs R., Tranel D., Cushman F., Hauser M. & Damasio A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908-911.

Königs, P. (2018). On the normative insignificance of neuroscience and dual-process theory. *Neuroethics*, *11*(2), 195-209.

Kohlberg L. (1971). From is to ought: How to commit the naturalistic fallacy and get away with it in the study of moral development. In Mischel T. (eds.), *Cognitive development and epistemology*, pp. 151-235. New York: Academic Press.

LaBar K. S., Gitelman D. R., Parrish T. B., Kim Y. H., Nobre A. C. & Mesulam M. (2001). Hunger selectively modulates corticolimbic activation to food stimuli in humans. *Behavioral neuroscience*, *115*(2), 493.

Lim S. L., Padmala S. & Pessoa L. (2009). Segregating the significant from the mundane on a moment-to-moment basis via direct and indirect amygdala contributions. *Proceedings of the National Academy of Sciences*, *106*(39), 16841-16846.

Maess B., Koelsch S., Gunter T. C. & Friederici A. D. (2001). Musical syntax is processed in Broca's area: an MEG study. *Nature neuroscience*, *4*(5), 540-545.

Maiese M. (2014). Moral cognition, affect, and psychopathy. *Philosophical Psychology*, *27*(6), 807-828.

Málková L., Gaffan D., & Murray E. A. (1997). Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *Journal of Neuroscience*, *17*(15), 6011-6020.

Martin A., Haxby J. V., Lalonde F. M., Wiggs C. L. & Ungerleider L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science*, *270*(5233), 102-105.

Martin A., Wiggs C. L., Ungerleider L. G. & Haxby J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, *379*(6566), 649-652.

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.

Meteyard L., Cuadrado S. R., Bahrami B. & Vigliocco G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804.

Mitchell M. (2006). Complex systems: Network thinking. *Artificial intelligence*, *170*(18), 1194-1212.

Moll J. & de Oliveira-Souza R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in cognitive sciences*, *11*(8), 319-321.

Moore G. E. (1903/1959). *Principia Ethica.* Cambridge: Cambridge University Press.

Moretto, G., Làdavas, E., Mattioli, F., & Di Pellegrino, G. (2009). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of cognitive neuroscience*, *22*(8), 1888-1899.

Moruzzi G. & Magoun H. W. (1949). Brain stem reticular formation and activation of the EEG. *Electroencephalography and clinical neurophysiology*, *1*(1-4), 455-473.

Nagel J. & Waldmann M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 237.

Nichols S. (2021). *Rational rules: Towards a theory of moral learning*. Oxford University Press.

Nisbett R. E. & Wilson T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84*, 231-259.

Nishitani N., Schurmann M., Amunts K. & Hari R. (2005). Broca's region: from action to language. *Physiology*, *20*(1), 60-69.

Öhman A. & Mineka S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, *108*(3), 483.

Paradiso S., Johnson D. L., Andreasen N. C., O'Leary D. S., Watkins G. L., Boles Ponto L. L. & Hichwa R. D. (1999). Cerebral blood flow changes associated with attribution of emotional valence to pleasant, unpleasant, and neutral visual stimuli in a PET study of normal subjects. *American Journal of Psychiatry*, *156*(10), 1618-1629.

Pessoa L. (2008). On the relationship between emotion and cognition. *Nature reviews neuroscience*, *9*(2), 148-158.

Phillips M. L., Medford N., Young A. W., Williams L., Williams S. C. R., Bullmore E. T., Gray J. A. & Brammer M. J. (2001). Time courses of left and right amygdalar responses to fearful facial expressions. *Human brain mapping*, *12*(4), 193-202

Pillutla M. M. & Murnighan J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational behavior and human decision processes*, *68*(3), 208-224.

Poldrack R. A. (2006). Can cognitive processes be inferred from neuroimaging data?. *Trends in cognitive sciences*, *10*(2), 59-63.

Pollard B. (2005). The rationality of habitual actions. *Proceedings of the Durham-Bergen Philosophy Conference*, *1*, 39–50.

Prinz J. (2006). The emotional basis of moral judgments. *Philosophical explorations*, *9*(1), 29-43.

Raia R. (2023). An analysis of conceptual ambiguities in the debate on the format of concepts. *Phenomenology and the Cognitive Sciences*, 1-26.

Railton P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics*, *124*(4), 813-859.

Redouté J., Stoléru S., Grégoire M. C., Costes N., Cinotti L., Lavenne F., Le Bars D., Forest M. G. & Pujol J. F. (2000). Brain processing of visual sexual stimuli in human males. *Human brain mapping*, *11*(3), 162-177.

Rosas A. & Aguilar-Pardo D. (2020). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning*, *26*(4), 534-551.

Sander D., Grafman J. & Zalla T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences*, *14*(4), 303-316.

Sauer H. (2012a). Educated intuitions. Automaticity and rationality in moral judgement. *Philosophical Explorations*, *15*(3), 255-275.

Sauer H. (2012b). Morally irrelevant factors: What's left of the dual process-model of moral cognition?. *Philosophical Psychology*, *25*(6), 783-811.

Sauer H. (2017). *Moral judgments as educated intuitions*. MIT press.

Saunders L. F. (2016). Reason and emotion, not reason or emotion in moral judgment. *Philosophical Explorations*, *19*(3), 252-267.

Schwarz N. (2002). Feelings as information: moods influence judgment and processing strategies. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 534–547). New York: Cambridge University Press.

Schwarz N. & Clore G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of personality and social psychology*, *45*(3), 513.

Shiv B. & Fedorikhin A. (2002). Spontaneous versus controlled influences of stimulus-based affect on choice behavior. *Organizational Behavior and Human Decision Processes*, *87*(2), 342-370.

Simmons W. K., Martin A. & Barsalou L. W. (2005). Pictures of appetizing foods activate gustatory cortices for taste and reward. *Cerebral cortex*, *15*(10), 1602-1608.

Singer P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, *1*(3), 229-243

Smith A. (1759/1976). *The theory of moral sentiments*. Oxford: Oxford University Press.

Sober E. & Wilson D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, Mass.: Harvard University Press.

Sinclair, R. C., Mark, M. M., & Clore, G. L. (1994). Mood-related persuasion depends on (mis) attributions. Social Cognition, 12(4), 309-326.

Stichter, M. (2018). *The skillfulness of virtue: Improving our moral and epistemic lives*. Cambridge University Press.

Trémolière B. & Bonnefon J. F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*(7), 923-930.

Trivers R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology 46*, 35-57.

van Honk J., Eisenegger C., Terburg D., Stein D. J. & Morgan B. (2013). Generous economic investments after basolateral amygdala damage. *Proceedings of the National Academy of Sciences*, *110*(7), 2506-2510.

Winstanley C. A., Theobald D. E., Cardinal R. N. & Robbins T. W. (2004). Contrasting roles of basolateral amygdala and orbitofrontal cortex in impulsive choice. *Journal of Neuroscience*, *24*(20), 4718-4722.

Winston J. S., Strange B. A., O'Doherty J., & Dolan R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces, *Nat. Neurosc. 5*, 277-283.

Wrangham R. & Peterson D. (1996). *Demonic males: Apes and the origins of human violence*. Boston: Houghton Mifflin.

Young A. W., Aggleton J. P., Hellawell D. J., Johnson M., Broks P. & Hanley J. R. (1995). Face processing impairments after amygdalotomy. *Brain*, *118*(1), 15-24.

Zalla T., Koechlin E., Pietrini P., Basso G., Aquino P., Sirigu A. & Grafman J. (2000). Differential amygdala responses to winning and losing: a functional magnetic resonance imaging study in humans. *European journal of Neuroscience*, *12*(5), 1764-1770.

Zerilli J. (2019). Neural reuse and the modularity of mind: where to next for modularity?. *Biological Theory*, *14*(1), 1-20.