Experimental moral psychology and its normative implications

Bruni, Tommaso

Experimental moral psychology and its normative implications

Tommaso Bruni

PhD in "Foundations Of The Life Sciences And Their Ethical Consequences" (FOLSATEC)
University of Milan – Italy
University Registration Number: R08432

Doctorat ès Lettres – Philosophie
Université de Genève – Suisse
Etudiant No. 09-313-941

Date of submission: December 20th, 2012

Supervisors:
Prof. Giovanni BONIOLO – University of Milan
Prof. Bernardino FANTINI – Université de Genève
Prof. Patrik VUILLEUMIER – Université de Genève
Bernard BAERTSCHI, M.E.R. – Université de Genève

If you are distressed by anything external, the pain is not due to the thing itself, but to your estimate of it; and this you have the power to revoke at any moment.

Marcus Aurelius, *Meditations*

Index

Abbreviations

ACC: Anterior Cingulate Cortex

AHA: *Ad Hominem* Argument

AMIF: Argument from Morally Irrelevant Factors

ATL: Anterior Temporal Lobe

BOLD: Blood Oxygen Level Dependent

CDA: Cultural Debunking Argument

CMJ: Considered Moral Judgment

CRT: Cognitive Reflection Test

DDA: Doctrine of Doing and Allowing

DDE: Doctrine of Double Effect

DLPFC: Dorso-Lateral Prefrontal Cortex

DMPFC: Dorso-Medial Prefrontal Cortex

EDA: Evolutionary Debunking Argument

EEA: Environment of Evolutionary Adaptiveness

EFEC: Event-Feature-Emotion Complexes

EMI: Epistemological Moral Intuitionism

fMRI: functional Magnetic Resonance Imaging

FTD: Fronto-Temporal Dementia

IAT: Implicit Association Task

IPL: Inferior Parietal Lobule

MFT: Moral Foundations Theory

MPFC: Medial Prefrontal Cortex

PCC: Posterior Cingulate Cortex

PET: Positron Emission Tomography

PFC: Prefrontal Cortex

PLF: Posterior Lateral Fusiform

p-STS: posterior Superior Temporal Sulcus

RCT: Rational Choice Theory

RE: Reflective Equilibrium

rTMS: repeated Transcranial Magnetic Stimulation

SCR: Skin Conductance Response

SIM: Social Intuitionist Model

TMS: Transcranial Magnetic Stimulation

TPJ: Temporal Parietal Junction

TPR: Total Peripheral Resistance

UG: Ultimatum Game

UMG: Universal Moral Grammar

UT: Unconscious Thought

VMPFC: Ventro-Medial Prefrontal Cortex

WRE: Wide Reflective Equilibrium

Figures Index

Abstract

This thesis explores the relationships between experimental moral psychology and normative ethics. It specifically examines neuromoral theories, according to which a deeper understanding of the machinery for moral judgments could lead humans to make *better* moral judgments. In this thesis I use the widely discussed neuromoral theory by Joshua Greene as a case study. I first examine Greene's descriptive claims in experimental moral psychology (Ch. 2). Then I review descriptive hypotheses concerning human moral cognition that are alternative to Greene's and I conclude that the data available so far are not sufficient to rule all alternatives out. Theories in experimental moral psychology are presently underdetermined by the data. In Ch. 4 I critically delve into Greene's neuromoral theory, highlighting its problematic points. Greene derives normative consequences from empirical results through the Argument from Morally Irrelevant Factors. This argument is not persuasive because it is not backed by an analysis of judgments about moral relevance of factors, such as "Spatial distance is a moral irrelevant factor", which are key premises in Greene's argument. I argue that these judgments cannot be taken for granted because they are often deeply controversial. Greene also falls in a recurring problem, i.e. the so-called 'meta-normativity problem'. It is not clear what kind of normativity neuromoral theorists are referring to when they say that empirical science could help humans make *better* moral judgments. These shortcomings make Greene's neuromoral theory unconvincing. However, Greene's descriptive work has greatly contributed to further the understanding of the machinery for moral judgments.

Introduction

This dissertation explores the relationships between experimental moral psychology and normative ethics. Experimental moral psychology is the empirical study of human moral behavior, including the making of moral judgments, and is carried out with both the methods of experimental psychology (reaction times, eye-tracking, skin conductance response, and so on) and neuroscientific methods (functional Magnetic Resonance Imaging – henceforth fMRI, electroencephalography, etc). Experimental moral psychology describes the machinery for moral judgments. This machinery underpins the formation and processing of moral judgments, where "moral judgment" is taken to refer to a kind of mental states rather than a kind of linguistic utterances. "Processing" comprises all the ways in which moral judgments are used within the mind after their formation, including the production of overt behavior. In moral judgments properties such as "morally mandatory", "morally forbidden", "morally praiseworthy", "morally good", etc. are attributed to an action carried out by a competent human being or to her character. There are similar normative concepts that originate from the aesthetic, economic, political, legal, prudential domains, and they give rise to mental states that are similar to moral judgments, i.e. political, economic, etc. judgments. The difference between non-moral normative judgments and moral judgments is an extremely interesting and debated issue, but I mention it just to set it aside. In what follows, I take for granted that there exists a moral domain that is significantly different from other forms of normativity, a domain to which moral judgments correspond[1]. Another very interesting issue amounts to whether this moral domain is almost fixed across cultural groups and historical periods, i.e. over space and time, or it undergoes substantial variations along these dimensions. Of course this issue is a matter of degree and, as all matters of degree, involves some vagueness. Again, I

[1] This assumption is in turn contested. I briefly discuss in § 3.6. what kind of charges have been addressed to it.

mention this just to put it aside. Even though I broach this problem at some point, I do not deal with this very interesting question in any extensive or systematic way.

This dissertation tries to answer the question whether results in experimental moral psychology can have normative consequences, i.e. whether they can tell a Western audience which, among Western normative ethical theories (e.g. consequentialism, deontology, and virtue ethics in their multifarious flavors), they ought to follow. A lively debate about these issues started in 2001 due to the publication of two landmark papers (Greene et al. 2001; Haidt 2001) and has not stopped since. The debate is partially a re-installment of an older debate about the role of evolutionary biology in ethics (cf. for instance Kitcher 2006/1993; Ruse 1986, Ruse and Wilson 2006/1986, Singer 1981; Wilson 1975, 1979), but the older debate focused much more on meta-ethics than on normative ethics.

Current empirical results in empirical moral psychology could have important consequences for meta-ethics too (whether this is the case or not is of course a moot point) but in this dissertation I deal with consequences for normative ethics only and I set aside issues that have to do with meta-ethics. As the debate about the relationship between empirical moral psychology and normative ethics has been monopolized in the last years by Greene's (2008a) very bold claims, I focus on Greene's descriptive and normative views.

I make a series of claims.

First, I maintain that there are many competing models in experimental moral psychology and that at present Greene's model does not stand out of the fray as the uncontested winner. More data points are needed to decide which theory (if any) is the correct one.

Secondly, I claim that, even assuming *arguendo* that Greene's depiction of the machinery for moral judgments is correct, his main normative claim does not follow.

Thirdly, Greene mostly uses the Argument from Morally Irrelevant Factors (henceforth AMIF) to derive normative conclusions from scientific data. This argument could work,

even though it features some problems. However, for it to run correctly, uncontroversial normative premises about the moral relevance of factors are needed. But since such premises can be deeply controversial, Greene cannot use this strategy to buttress the normative consequences he holds dear.

Here is a short overview of the dissertation.

In Chapter 1 I briefly examine the concept of moral intuition, since it is fundamental to address most descriptive models in experimental moral psychology, including Greene's. So it is necessary to understand exactly what people in the debate are talking about when they use this expression.

In Chapter 2 I examine Greene's descriptive theory, reviewing the supporting evidence and discussing the main criticisms that have been made.

In Chapter 3 I survey some important theories in experimental moral psychology that are alternative to Greene's view and I show how the empirical evidence on which Greene's descriptive theory is based is not sufficient to rule out these alternative hypotheses. At the same time, I survey some normative conclusions that have been drawn from some of these descriptive models and show that they suffer from recurring problems.

In Chapter 4, I deal with Greene's claim that an improved knowledge of the machinery for moral judgments ought to lead humans to reject deontology as a normative ethical theory, if not always at least under many circumstances. The problems that cripple neuromoral theories concern Greene's theory too. Furthermore, Greene's theory has specific issues that need to be examined.

In a final appendix, I examine a methodological problem concerning most of experimental moral psychology, i.e. idiosyncratic sampling, and I discuss ways to address this problem in future empirical research.

Chapter 1: Moral Intuitions

A full treatment of the concept of moral intuition, as well as a thorough discussion of the role of moral intuitions in moral thought, would deserve a dissertation in its own right. Therefore, I examine this concept here only insofar as this is required to discuss the relationship between experimental moral psychology and normative ethics.

The expression "moral intuition" has at least three meanings.

The categorization I am drawing is likely not to be exhaustive – it is possible that some other moral claims are called 'moral intuitions' by some authors and do not fit into the three categories I am going to sketch. However, this categorization will serve my purposes well: please take it as a working hypothesis.

First, there is the concept of moral intuitions as it is mostly used in experimental moral psychology. The most famous definition of the psychological concept of moral intuition is found in Haidt's (2001) landmark paper. Haidt defines a moral intuition as "the sudden appearance in consciousness of a moral judgment, including an affective valence […] without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion" (Haidt 2001, 818)[2]. Among philosophers, Sinnott-Armstrong (2008a, 47) similarly describes a moral intuition as "a strong immediate moral belief." "Strong" indicates that the believer will not give away the judgment easily. "Immediate" indicates the absence of inference and conscious processing, as in Haidt's definition. A good example of a moral intuition in the sense of Haidt and Sinnott-Armstrong is the moral judgment: "It is morally wrong that Mark and Julie, who are siblings, have sex with each other." Following this definition, an instinctive response to a

[2] Moral judgments are in turn defined by Haidt as "evaluations […] of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture" (2001, 817). Haidt leaves this definition broad on purpose. Haidt (2012, 270) gives an idea of his overall conception of morality by defining 'moral systems' in the following way: "Interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate self-interest and make cooperative societies possible." Therefore, Haidt describes a very wide moral domain.

moral case can be considered as a moral intuition. Although it was Haidt who lately made this concept popular, it can also be traced back to a long-standing tradition in philosophy. The concept of perceptual moral intuitions as "immediate judgment as to what has to be done or aimed at" is already present in Sidgwick (1907, 97). Sidgwick correctly noticed that an individual's moral intuitions about a single case can be inconsistent over time and that moral intuitions of different people concerning the same case often diverge. Sidgwick distinguished three types of moral intuitions: perceptual, dogmatic, and philosophical. Haidtian moral intuitions roughly correspond to perceptual intuitions. I will deal later with Sidgwick's philosophical intuitions, which represent the third meaning of "moral intuition" I avail myself of.

Haidtian moral intuitions have some specific features. First, they are always about specific cases and are distinct from moral principles precisely because they are not general. Secondly, moral intuitions are not reflective. These judgments are immediate, in the sense that they are caused by a cognitive mechanism that is roughly akin to perception. I do not want to enter all interpretations of the concept of 'intuition' in philosophy (for a useful review, see Pust 2012). It is sufficient to my purposes to clarify that these judgments are incompatible with reflection. Reflection can be used in various ways. When one reflects, one can think carefully about the features of the given case, or can carry out inferences. However, this concept of intuition excludes all of this and these judgments are not only non-inferential. They are more generally unreflective[3]. Thirdly, these judgments are related to emotions, even if it is not so clear how. In the original formulation by Haidt (2001), the emotional component of this kind of moral intuitions was pretty important and moral intuitions needed to have an affective part. The emotions involved could be multifarious: possibly disgust for the incest taboo, empathy for pain in "It is morally wrong to torture this terrorist," and so on. This has changed over time and Haidt now focuses more on immediateness than on emotion, but some non-necessary connection with emotion

[3] I assume reflection to be a necessary condition for inference.

remains. In other words, pangs of emotion (either negative or positive in valence) correlate with these judgments most of the times, although there is no necessary link.

Due to these features, the set of moral judgments that is identified by Haidt's and Sinnott-Armstrong's definition is quite broad: all moral judgments about cases that do not arise from conscious processing and are unreflective responses to some morally relevant[4] event in the world can be considered as "moral intuitions" in this sense. Many (but not all) moral evaluations that are given by participants in moral psychology experiments behind the solid walls of the lab count as moral intuitions in this sense. It is also possible to survey moral intuitions in the field, via questionnaires, or through the World Wide Web, using tools such as the Amazon Mechanical Turk[5] (cf. Paolacci, Chandler, and Ipeirotis 2010), provided that participants are cued to give fast judgments and are effectively stopped from reflecting.

Secondly, "moral intuition" can indicate the concept of "considered moral judgment" that has been used by Rawls and Daniels in the method of reflective equilibrium (henceforth RE) (Daniels 1979, 1980a, 1980b, 2011; Rawls 1951, 1999/1971). Please notice that I am referring to moral judgments as they are *before* going 'back and forth' in RE. Hence, "moral intuitions" in this sense are part of the *input* and not part of the *output* of RE. Whether "considered moral judgments" (henceforth CMJ) are the input or the output of RE is debatable as the interpretation of the Rawlsian text is quite complex. I stick here though to the interpretation of Daniels (2011), according to which CMJs enter together with moral principles and background theories in Wide Reflective Equilibrium (henceforth WRE) to yield justified judgments, principles, and theories. If "moral intuition" is thus interpreted, then a moral judgment must satisfy a long series of conditions to be seen as a "moral intuition," enter RE, and interact with moral principles (and, depending on the version of RE, background theories). In particular, the authors of these judgments must not be in

[4] I now leave aside the very important and interesting issue concerning what events, actions, or personality traits count as morally relevant.

[5] https://www.mturk.com/mturk/welcome. Accessed August 29th, 2012.

contexts that generate frequent errors, must be aware of the biases created by their preferences, must possess qualities such as intelligence, empathy, and reasonableness, and must have access to information about the world, the case at issue, and the interests in conflict (cf. Rawls 1951). Hence a simple gut reaction to a situation does *not* qualify as a moral intuition according to this standard. People are rarely under the correct conditions for making moral intuitions in the Rawlsian sense. Out of the very large number of moral judgments people make every day, just a few are CMJs. This means that these judgments are extremely rare and are unlikely to be recorded in survey studies done outside the lab. They could be rare even inside the lab, given the strident demands in terms of cognitive ability and level of information that are put on the person that makes them. This concept of "moral intuition" is so idealized that theorists sympathetic with RE (for instance Van Thiel and Van Delden 2010) have proposed to substitute it with the Haidtian concept in order to allow moral intuitions gathered via empirical research into RE. In order to better distinguish this concept from the previous one, I will make use of "CMJ" to refer to the Rawlsian concept.

There is then a third concept, that I label as "rational moral intuitions". They derive from Sidgwick's third kind of intuitions, i.e. philosophical intuitions. According to Sidgwick, there are three philosophical intuitions, i.e. self-evident moral axioms. These axioms are too general for normative conclusions about cases to be deduced from them[6], but at least they are not tautological, as in Sidgwick's opinion many moral statements that purport to be self-evident are. The three ethical intuitions are the axiom of justice[7], the axiom of prudence[8], and the axiom of benevolence[9]. These three propositions are just examples of

[6] "There are certain absolute practical principles, the truth of which, when they are explicitly stated, is manifest. But they are of too abstract a nature, and too universal in their scope, to enable us to ascertain by immediate application of them what we ought to do in any particular case." Sidgwick (1907, 379)

[7] "It cannot be right for A to treat B in a manner in which it would be wrong for B to treat A, merely on the ground that they are two different individuals, and without there being any difference between the natures or circumstances of the two which can be stated as a reasonable ground for difference of treatment." Sidgwick (1907, 380).

[8] "The mere difference of priority and posteriority in time is not a reasonable ground for having more regard to the consciousness of one moment that [*sic*] to that of another." Sidgwick (1907, 381).

rational moral intuitions – others can be imagined, even though according to Sidgwick there are just three of them. Rational moral intuitions are general, are not about cases, and have a high level of plausibility. Another example could be "It is morally preferable to save more than one human life from death than just one" (cf. Singer 2005). These rational intuitions are interesting because consequentialist theorists often have to employ them even though they reject all other moral intuitions (cf. Singer 1974). As Sinnott-Armstrong (2008a) correctly notes, all theorizing in normative ethics has to start from some normative claims.

[9] "The good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other; unless, that is, there are special grounds for believing that more good is likely to be realized in the one case than in the other." Sidgwick (1907, 382).

Chapter 2: Greene's Descriptive View

2.1. Greene's dual-process model

Greene's main idea is that the machinery for moral judgments is not unitary, but divided in two separate circuits. According to his view, these two systems have different ways of working, i.e. they process information in different ways, and they have different outputs, i.e. they make different kinds of judgments (Greene et al. 2001, 2004, 2008, 2009). In characterizing the two systems, Greene draws on a well-established tradition in experimental psychology: dual-process theories have been employed to explain a wide gamut of cognitive phenomena. The most notable theorist that has made use of dual-process models is Nobel-laureate Daniel Kahneman (2003, 2012), but these models have also been used by a wide range of scientists in other domains of psychology, e.g. in social psychology (cf. Chaiken and Trope 1999). Therefore, in Greene's model there are two systems that are in charge with moral judgments. The first system is usually called System1[10]. Its operations are automatic, fast, inflexible, cognitive impenetrable (the subject is normally not aware of them), effortless, associative, and emotionally charged. The second system is known as System2[11]. In contrast with System1, its operations are conscious, potentially controlled, slow, flexible, cognitive penetrable (the subject is aware of them), effortful, and serial. System1 is evolutionarily older than System2 and is based on emotional reactions to particular actions. If the reaction is a negative one, the resulting moral judgment will be a negative one too, i.e. it will be a judgment that the action ought not to be done. If the emotional reaction is a positive one, the resulting moral judgment will be positive, i.e. the judgment will be that the action should be carried out. According to Greene, System1 reacts with negative pangs of emotion and ensuing moral

[10] Using an analogy with a camera, Greene also calls it the automatic mode.
[11] Drawing on the same analogy as in Footnote 9, Greene also calls this system the manual mode.

condemnation to particular features of scenarios, such as violence carried out by means of personal force. Personal force is thus defined by Greene and co-workers:

> 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. Thus, applications of personal force, so defined, cannot be mediated by mechanisms that respond to the agent's muscular force by releasing or generating a different kind of force and applying it to the other person. (Greene et al. 2009, 365).

A cognitive system that reacted to personal force limited inter-personal violence in the Environment of Evolutionary Adaptiveness[12] (henceforth EEA) and was hence fitness-enhancing in a hyper-social mammal species such as *Homo sapiens*, in which some form of group selection is likely to have occurred and to occur now[13]. System1 is likely to react to other features of scenarios too, such as violations of fairness[14], the intention of bringing

[12] I borrow this expression from Savulescu and Sandberg (2008). The EEA roughly corresponds to the Pleistocene hunter–gatherer existence.

[13] The issue whether group selection really occurs in nature has been a topic of dispute for decades. There is no room to consider the arguments for the different positions, so that I will just describe how the consensus view changed over time. Group selection was commonly accepted in the evolutionary biology community until the 1960s. Then the influential critique by G. C. Williams (1966) changed the landscape. Altruistic behaviors in metazoans were then increasingly explained through kin selection and reciprocal altruism. These two mechanisms managed to account for many phenomena, but, however important, they were unable to explain the whole gamut of altruistic behaviors in humans. It is difficult, for example, to explain altruistic punishment (A punishes B because B has betrayed C's trust or has violated an established rule of the human group to which A, B, and C belong) using kin selection and reciprocal altruism only. These difficulties have lead Sober and Wilson (1999) to re-introduce group selection, at least under certain conditions. To my knowledge, the position advocated by Sober and Wilson counts nowadays as the consensus in evolutionary biology.

[14] Cf. for example the hemodynamic bilateral insular activation reported by Hsu, Anen, and Quarz (2008). This activation takes place when a participant perceives that goods have been distributed in an unequal way and acts in order to enforce equality. Concerning fairness, an interesting case is underscored by Dean (2010). It involves the moral dilemma called *Lost Wallet*, in Greene et al. (2001), Supplementary Material. In this dilemma you find the wallet of a wealthy person, who does not need the several hundred dollars present in the wallet itself. On the contrary, you have been recently hit by hard times and that money would help you and your family a lot. So, from the point of view of act consequentialism, keeping the money maximizes aggregate well-being, so that you ought to keep the money. However, most people give deontological responses to this dilemma and send the wallet back without touching the money. This seems to fit poorly with Greene's model, because this is an impersonal dilemma whose emotional salience is close to none. There seems to be a dissociation between emotion and deontological judgments. Hence, what is triggering our alleged System 1 here? The BOLD activation reported by Hsu and co-workers helps Greene defuse this objection. Greene could assume that there are automatic settings that react to violations of fairness and that do not require high levels of emotional salience in the stimuli to be activated. Empirical research on this is still at the onset, but it is sufficient to allow the dual-process model to respond to this objection.

about physical or psychological harm (Greene et al. 2009), or disgust[15]. The moral System1 works well in most moral situations of everyday life. Greene compares System1 to the automatic mode of a digital camera (Greene 2010): it is highly efficient but not flexible. In contrast, System2 is much more flexible, but it is much slower and less efficient. It carries out a cost-benefit analysis of the scenario and decides accordingly, roughly following the dictates of rational choice theory (henceforth RCT)[16]. In particular, System2 considers different possible actions. Then the consequences that such actions are likely to produce are identified, weighted for their probabilities, and compared. If this comparison arrives at the conclusion that a specific action is likely to bring about bad consequences – or consequences that are worse than those brought about by the alternatives – then a judgment that the action ought not to be done is formed. *Mutatis mutandis* the same holds for actions whose expected consequences are good and moral judgments of approbation.

The two systems allegedly have different neural correlates. Greene tried to uncover the neural correlates of the two moral systems in his two fMRI papers (Greene et al. 2001, 2004). He identified the middle frontal gyrus, the angular gyrus, the posterior cingulate cortex (henceforth PCC), parts of the superior temporal sulcus, and the ventromedial prefrontal cortex (henceforth VMPFC) as the neural correlates of System1. The dorsolateral prefrontal cortex (henceforth DLPFC) and the Inferior Parietal Lobule (henceforth IPL) are instead the main neural correlates of System2. Greene claims that the areas that correspond to System1 are areas linked to emotion and the areas that correspond to System2 are areas linked to reasoning. Partitioning the brain in emotional and rational areas is however complex and perhaps unfeasible (cf. Pessoa 2008).

[15] As in the case of consensual incest. Cf. Schaich Borg et al. (2008) for the (rather remarkable) neural correlates of incest-related disgust.

[16] RCT is an application of the mathematical theory of games and it constitutes the groundwork of most of current micro-economics. It stems from the seminal work by Von Neumann and Morgenstern (1944). In RCT the rational consumer is conceived as an utility maximizer. However, there is an important difference between RCT and moral System2. An agent driven by RCT is usually self-interested, whereas an agent driven by moral System2 is not necessarily so. In contrast, an agent that follows System 2 is often committed to impartiality, in the sense that for the impartial agent the good of any one individual is of no more importance than the good of any other (including herself).

By the way, it is not very clear what "emotion" means in Greene's theory. Both emotion and cognition are in Greene's opinion forms of cognition and in some texts, such as Greene (2008a, 40), we find a difference between narrow cognition and cognition in general. Narrow cognition equates with reasoning, whereas cognition in general refers to all forms of information processing and therefore includes emotion if this is understood as a form of stimulus-dependent processing[17]. A discussion of theories of emotions would be both unnecessary and unfeasible in a text like the present one. This is especially true because Greene never mentions the three theories that have been running the emotion debate in psychology in the last decades: basic emotions theory (cf. Ekman 1992), appraisal theory (among others, cf. Scherer 2009a, 2009b), and constructivism (cf. Russell 2003). Greene and co-workers take 'emotion' to indicate "representations that have direct motivational force" and 'reasoning' to indicate the processing of representations that have "no direct motivational force of their own" (Greene et al. 2004, 397-398). In addition to the exact nature of emotion in the model, it is also unclear how tight the emotion–System1 connection is. In particular, Greene (2008a) interestingly introduces a distinction between two kinds of emotions: alarm-like emotions, that drive the automatic, blind responses of System1, and currency-like emotions, that are used as a means of comparison of different goods in cost-benefit analysis. Hence, currency-like emotions would be necessary to the correct working of System2. Shenhav and Greene (2010) suggest, with a strong backing from empirical literature, that the VMPFC is not simply an emotional area, but an area that has to do with the representation of value and reward and in which different emotions get integrated in order to yield an overall representation of value. They conjecture that currency-like emotions could be processed by the VMPFC. However, even if System2 might process some emotional component in order to create moral judgments, this does not stop the DLPFC and parietal areas to be the neural basis of a control network whose main

[17] There are important theories of emotions, such as Russell's (2003) constructivism, according to which emotion has little to do with stimuli and much more to do with the continuous categorization of a mental state known as 'core affect', whose two dimensions are arousal and pleasantness.

function is to inhibit prepotent emotional responses, as Cushman and colleagues (2011) clearly state. In other words, System2 cannot be taken to be 'emotional' as long as it is identified with a function of the DLPFC and a set of parietal regions. Even though Greene cannot at the moment spell out the exact role of emotion in his model and considers the integration of reasoning and emotion in decision making as an important topic for future research (Cushman et al. 2011), the distinction between narrow cognition and emotion in his model is pretty strong, as traditionally in dual-process theories of mental functioning.

Leaving aside the topic of emotions, the two systems produce the same judgment most of the time. Actions that trigger the moral System1 are usually condemned by the moral System2 too, as they are likely to bring about more harms than benefits. But there are some cases in which the two systems evaluate a given scenario in different ways. One of these situations is the *Footbridge* version of the notorious Trolley dilemma. Consider these two variants of the dilemma.

In the *Switch* scenario, one can hit a switch that will turn a runaway trolley away from five people and onto a track where another individual is located. As a result of this, the individual located in this other track will be killed but the five people – which would otherwise have died – will be saved.

In the aforementioned *Footbridge* scenario, one can push a person (a bystander wearing a huge, heavy backpack) off a footbridge and into the path of a runaway trolley. This will result in the death of the bystander, but the lives of five individuals further down the track – which would otherwise have died – will be saved because the body and the backpack are heavy enough to deflect the course of the trolley.

People react in very different ways to the two scenarios: cross-culturally, 89% of experimental participants judge that it is permissible to hit the switch, but only 11% judge that it is permissible to push the bystander (Hauser, Young, and Cushman 2008). These are not the only variants of the trolley dilemma: there are several others. The whole gamut of trolley cases have been one of the favorite topics in ethics in the last fifty years (Foot

1978/1967; Thomson 1976, 1985, 2008). The problem with these hypothetical scenarios is that people respond to them in very different ways even though they seem to be very similar from a philosophical point of view. In both *Switch* and *Footbridge* the agent can either kill one person to save five or let the five die, but the response is not the same even though the scenarios are apparently similar. Hence, philosophers have tried to spot the relevant differences between the scenarios in order to justify the divergence in people's intuitive responses, but each explanation that has been devised so far has been unfortunately countered by some counter-example, so that trolley dilemmas are as puzzling to philosophers today as they were fifty years ago.

Going back to Greene, the *Footbridge* case is an alleged instance of conflict between System1 and System2. System1 condemns pushing the bystander down to her death as a violent and illegitimate harm, whereas System2 favors the option of pushing the man because more human lives would thus be saved. From the descriptive point of view, System1 regularly prevails in case of conflict, even though System2 can sometimes (11% of cases) override the prepotent force of the emotional response and enforce the result of a cost-benefit analysis. However, when System1 is not engaged because the killing is not carried out by physical force, i.e. in *Switch*, System2 takes control and recommends the action that leads to the maximization of overall benefit among all people involved. Another important distinction in Greene's theory is that between characteristically deontological judgments (henceforth deontological judgments) and characteristically utilitarian judgments (henceforth utilitarian judgments). Deontological judgments are moral judgments that are easy to justify in deontological terms and difficult to justify in consequentialist terms, such as the judgment that it is morally impermissible to kill one person in order to save five. This does not necessarily mean that they are carried out for deontological reasons or by experimental participants that subscribe to deontology. Conversely, utilitarian judgments are easy to justify in consequentialist terms and difficult to justify in deontological terms, such as the judgments that it is morally permissible to kill

one in order to save five[18]. According to Greene's dual-process hypothesis, deontological judgments are typically the product of System1, while utilitarian judgments are typically the product of System2. Deontological judgments have the typical features of moral intuitions in the Haidtian sense, as they share the traits of System1 described above. In contrast, System2 needs some rational intuitions to work and any way it requires some account of well-being to maximize it[19]. From the mapping between deontological judgments and System1 on the one hand and between System2 and utilitarian judgments on the other hand, Greene derives his most controversial descriptive claim: deontology and utilitarianism[20] as substantive ethical theories count as "psychological natural kinds" (Greene 2008a, 36). So, not only the two systems generate judgments that are more easily justified by the one theory than by the other, but both theories themselves are attempts at justifying the operation of the systems. Hence, they exist *because* the systems exist and are the result of reflection on their output. This does not mean that the people that use System1 or System2 must have any necessary knowledge of these ethical theories. Those who judge that the bystander must be pushed down the bridge do not need to be card-carrying utilitarians (Greene, forthcoming, footnote 10 in § II). Humans can avail themselves of System2 without necessarily subscribing to a substantive ethical theory. As a matter of fact, the systems exist *before* the elaboration of normative ethical theories, which stem from the necessity of justifying their output in front of other human beings. However, these two substantive ethical theories are closely linked to the systems, so that the upholders of deontology are defending the judgments that result from System1 and the upholders of

[18] However, it must be considered that the consequentialist may ask us *not* to push the bystander down the bridge to her death. For example, a rule utilitarian might say that it makes sense to uphold an across-the-board prohibition against intentional killing of fellow humans because in the long run not allowing people to intentionally kill humans is more beneficial to society than allowing for exceptions in cases such as *Footbridge*. The freedom granted to individual agents of construing exceptions to this rule would lead to many instances of wrongful killing we do not want to occur. It is interesting to notice that this rule consequentialist argument also stops the agent to pull the switch in the homonymous dilemma, since it advocates a general ban against intentional killing and the intervention proposed in *Switch* amounts to killing a fellow human being.

[19] For this requirement, cf. de Lazari-Radek and Singer (2012, 27).

[20] I will refer to this theory both with "consequentialism" and "utilitarianism", although this is imprecise. Greene has in mind a form of maximizing act utilitarianism.

utilitarianism are defending the judgments that result from System2. In case of conflict between the two systems, in Greene's eyes deontologists claim that System1 should take precedence over System2 and, vice versa, utilitarians claim that System2 should take precedence over System1. As I argue in § 2.3, the empirical evidence for the 'psychological natural kinds' thesis is weak, precisely because Greene admits that, at least in many cases, judgments and theories can be uncoupled and experimental results tell something about judgments only.

2.2. The evidence for the dual-process view

What kind of empirical evidence are Greene's descriptive claims based upon? Some important piece of evidence comes from results obtained by Greene himself and coworkers, but Greene claims that the work of various labs all over the world contributes to corroborate the dual-process hypothesis. I briefly review the evidence here. One of the advantages of Greene's supporting evidence is its being domain-specific, i.e. specifically linked to morality, as it directly comes from empirical investigation about moral dilemmas, i.e. the so-called "trolleyology."

In Greene's opinion (2010, 28) the single strongest piece of evidence is the experiment carried out independently and roughly at the same time by Koenigs et al. (2007) and by Ciaramelli et al. (2007) on VMPFC patients. Lesions to the VMPFC are very famous in moral psychology, mostly due to two patients, 19th-century railway worker Phineas Gage (cf. Damasio 1994, ch.1 – *Unpleasantness in Vermont*) and patient EVR (Damasio 1994; Saver & Damasio 1991). Both showed preserved general cognitive activity, conserved capacity of understanding moral norms according to Kohlbergian standards[21], but

[21] Of course Kohlbergian standards did not exist at Gage's time, but Gage seemed to understand the content of the moral norms present in his society. Nevertheless, he often did not act upon them. For Kohlbergian standards, that assume that morality has mostly to do with reasoning and are relative to the justification of moral judgments, see Kohlberg (1969).

disastrous decision making in real life, both in moral and in non-moral situations[22]. The decision making deficit shown by these patients is multifaceted. They exhibit abnormal preference judgments (preferences tend to be inconsistent; Fellows and Farah 2007), difficulties in reversal learning[23] (Fellows and Farah 2003), and abnormal information acquisition patterns in multi-attribute decision making (Fellows 2006). Finally, studies by Shamay-Tsoory and colleagues (2003, 2009) show that VMPFC patients are impaired in affective Theory of Mind (i.e., attributing affective mental states to others). Even though the cognitive deficits of these patients are complex, they regularly show a condition known as "emotional blunting," i.e. most of their emotions are less intense than in the neuro-normal population. In the experiments by the groups of Koenigs and Ciaramelli, VMPFC patients were administered moral dilemmas from Greene et al. (2001) and made significantly more utilitarian judgments, such as "it is morally permissible to throw the bystander down the footbridge to save five," than healthy controls and non-frontal patients in high-conflict personal dilemmas, i.e. dilemmas that involve the use of personal force and that elicit a significant level of disagreement in the subject's responses[24]. This utilitarian bias in emotionally blunted patients would support Greene's idea that utilitarian judgments are driven by a cognitive system that regularly competes with emotions. If emotions are weaker, utilitarian judgments have the upper hand.

The VMPFC result has been contested, though, by Kahane and Shackel (2008) on the one hand and by Moll, De-Oliveira Souza, and co-workers (2007, 2008) on the other hand.

Kahane and Shackel underlined that the dilemmas used by Koenigs et al. (2007) have been selected on the basis of behavioral features of the subjects responses and not on the basis of content. They commented that the only factor that can make a response "utilitarian" is the

[22] Again, I assume *arguendo* that it is possible to identify the boundary between moral and non-moral domains in a vague but workable way.

[23] Reversal learning is the ability to change behavior when the pattern of rewards and punishments in a situation is reversed. If for instance an experimental participant gets money if she presses a red button (instead of a blue button) when an electric bulb switches on, and then she is told that from now on she has to press the blue button when the light is on, otherwise she would have to pay money back to the experimenter, she should change her behavior. Doing so counts as an instance of reversal learning.

[24] This in Koenigs et al. (2007). Ciaramelli et al. (2007) found the same effect, but for personal moral dilemmas, which is a somewhat less strict categorization.

content of the options proposed in the scenario. To buttress this claim, they asked professional moral philosophers to assess the dilemmas used by Koenigs et al. They report that these experts did not classify many of the "high-conflict" dilemmas as including an utilitarian option. Hence, Koenigs et al's results did not show that VMPFC patients had an utilitarian bias[25]. Nonetheless, Koenigs et al. (2008) re-analyzed their VMPFC patients' data using the classification proposed by Kahane and Shackel and found the same utilitarian bias they had spotted beforehand, so that this criticism doesn't seem to be particularly dangerous[26].

Moll et al. (2007; 2008, 168) remarked a discrepancy between this VMPFC patients result and another result by the same scientists (Koenigs and Tranel 2007). This experiment had to do with a very famous economic game, the Ultimatum Game (henceforth UG)[27]. Koenigs and Tranel showed that VMPFC patients rejected proposals more than usual. Since rejecting offers in the UG is normally seen as an irrational behavior, this seemed to show that VMPFC patients were less rational than healthy controls, a finding that apparently contradicted the interpretation given by Greene to the Koenigs et al. (2007) result[28]. More specifically, Koenigs and Tranel (2007) interpreted the UG behavior of VMPFC patients as a consequence of their alleged inability to regulate negative emotions. But VMPFC patients cannot be more rational and more irrational at the same time, Moll and colleagues argued. They suggested instead that the VMPFC patients suffered from a

[25] Since Ciaramelli et al. (2007) have used roughly the same dilemmas as Koenigs and co-workers, the criticism by Kahane and Shackel (2008) would apply to the Italian experiment too, even though Kahane and Shackel did not deal with the paper by the Italian research group.

[26] Greene correctly comments that "Kahane and Shackel proposed a more stringent test for the dual-process theory, and the dual-process theory passed with flying colors." (2010, 29)

[27] In the Ultimatum Game there are two players, a Proposer and a Responder. The Proposer is given a sum of money and must assign some part of it to the Responder. The determination of the exact amount, which can also be 0 or the whole sum, is left to the Proposer. The Responder can then accept or reject the Proposer's offer. If the Responder accepts, the money is actually divided between the two players according to the proposal. If the Responder rejects, the researcher takes the whole sum back and both players earn nothing. RCT dictates that Responders should always accept any positive offer, but this is not what happens in the lab. Responders tend to reject very unfair offers, such as 10% or 15% of the total lot (or less). This constitutes a form of costly punishment toward the Proposer, who was seen as stingy by the punishing Responder, as she offered too little.

[28] For a response by Greene to Moll's attack, see Greene (2007). However, Greene's reply is not very convincing because he could not avail himself of the decisive experiment by Moretti et al. (2008), which was published one year later.

specific dysregulation in prosocial sentiments that led them to make inappropriate utilitarian judgments, judgments that infringed upon well entrenched societal norms such as the prohibition towards intentionally killing fellow human beings. The interpretation by Koenigs and Tranel (2007) has been shown wrong, though. It is true that VMPFC patients rejected more in the UG than healthy controls, but not because they could not regulate their negative emotions. The key paper here is Moretti, Dragone, and Di Pellegrino (2008). Moretti and co-workers showed that VMPFC patients' rejection bias appeared just in specific environmental situations, i.e. when the monetary rewards were not immediately available in cash. If the game was played with cash in sight and the money was immediately distributed, no bias was found. This dooms Koenigs's and Tranel's interpretation of their own results. The correct interpretation follows the lines of Damasio (1994, ch. 9): VMPFC patient exhibit a specific 'myopia for the future', so that they cannot take into account consequences of their actions if they lie in the future and are not directly experienced at the time of the decision. Since both criticisms against the VMPFC result on moral dilemmas fail, this seems to count as evidence for Greene's model, provided that these patients are mainly seen as emotionally-blunted subjects. Nonetheless, as I have written above, the cognitive and affective deficits of these patients are extremely complex, so that it is not immediately clear *why* they make more utilitarian judgments than controls. However, Moretto and co-workers (2009) showed that VMPFC patients exhibited no Skin Conductance Response (henceforth SCR)[29] when they make utilitarian judgments in front of moral dilemmas involving personal force, contrary to healthy controls and non-frontal patients. SCR is quite tightly associated with emotion. So Greene has sufficient empirical backing to claim that the reason why VMPFC patients make more utilitarian judgments on high-conflict moral dilemmas is emotional blunting. Finally, the VMPFC evidence has been recently enriched by a study (Thomas, Croft, and Tranel 2011) showing that the

[29] A very slight increase in the degree of sweat production on the skin, that is regularly associated with emotional arousal. It is impossible to perceive it with the naked eye, but a pair of electrodes can measure skin conductance and detect that sweat production has increased, since sweat contains salt and salty solutions are good electricity conductors. Cf. Dawson, Schell, and Fillon (2007).

patients' utilitarian bias is even wider than previously thought. It also applies to dilemmas in which considerations of aggregate welfare are pitted against indirect violence to members of one's family (e.g. a version of *Switch* in which the one person on the side-track is the agent's daughter). Furthermore, the utilitarian bias generalizes to both the Self and the Other condition[30]. Summing up, the VMPFC evidence nicely dovetails Greene's dual-process model because it shows that utilitarian judgments are in competition with emotions.

Another similar piece of evidence is the study by Mendez, Anderson, and Shapira (2005). These researchers found a utilitarian bias on *Switch* and *Footbridge* in patients affected by Fronto-Temporal Dementia (henceforth FTD), another condition that brings about emotional blunting. However, the set of symptoms linked to FTD is very broad and to my knowledge no SCR measurement was made on these patients to show that emotional blunting was actually the cause of their utilitarian bias. Since this check is lacking, it is difficult to attribute the deficit to any particular facet of the complex cognitive deficit these patients exhibit, even though an analogical argument from VMPFC patients could be made.

An utilitarian bias has been recently attributed to another population that is known for its emotional blunting: low-anxiety psychopaths (Koenigs et al. 2011). Please notice that psychopaths in general do not show any utilitarian bias (Glenn et al. 2009b; Koenigs et al. 2011). Previous studies (Glenn et al. 2009a, 2009b) had shown that psychopaths in general show a reduced Blood Oxygen Level Dependent (henceforth BOLD) signal[31] in the

[30] The Self condition is the one in which the agent in the dilemma is the experimental participant that must make a moral judgment. The Other condition is that in which the agent in the dilemma is some different person from the experimental participant that must make a moral judgment. The distinction is equivalent to the one between first and third person judgments. People usually have a Self-Other bias (Nadelhoffer and Feltz 2008): they are more likely to approve of characteristically consequentialist judgments in the Other condition. VMPFC patients are not immune to this bias, but they tend to approve of these judgments more than controls in both conditions (i.e. Self and Other).

[31] The BOLD signal is what fMRI measures. It is a measure of cerebral blood oxygenation and flow. It correlates with metabolic neural activity. Cf. Logothetis (2008).

amygdala, a region that is usually connected with emotions[32], when confronted with personal, emotionally-loaded moral dilemmas. Glenn et al. (2009b) also report that psychopaths exhibit an increased activation in the DLPFC, the chief cognitive control area[33], when they read and judge about personal moral dilemmas. To sum up, most psychopaths have a hypo-emotional phenotype and a hyper-rational brain activation pattern, but low-anxiety psychopaths only show the utilitarian bias that is common among VMPFC patients. This seems to be good evidence for Greene's dual-process model, though.

In a very interesting study involving healthy participants, Koven (2011) measured two variables, 'Attention to Emotion' and 'Clarity of Emotion'. Then she checked whether these two measures had any correlation with the judgments made by participants in response to Greene's dilemmas. She found out that healthy participants with a high 'Clarity of Emotion', i.e. particularly good at understanding and distinguishing their emotions, tended to make significantly less utilitarian judgments. Drawing on work by Gohm (2003), Koven speculates that people that are particularly bad at grasping their moods and emotional states are very good mood regulators and emotion suppressors. They are normally overwhelmed by emotional states that they do not understand and hence systematically disregard emotions in decision making. People with a poor 'Clarity of Emotion' are likely to be among these "overwhelmed" participants, that are hence good at regulating emotions and ignoring their contents. So people that have bad 'Clarity of Emotion' make more utilitarian judgments and, on the other hand, people with high 'Clarity of Emotion' pass more non-consequentialist judgments, as they are apt at taking emotions into account and regularly use emotional information to make their decisions. If the conjecture held, it would nicely integrate in Greene's dual-process model.

[32] There are nonetheless alternative interpretations of the role of the bilateral amygdala in the brain. For an interpretation that sees the amygdala as a high connectivity hub, see Pessoa (2008, 152).
[33] To use Greene's words in a personal communication, "the accounting department."

Bartels (2008, 393) shows, in the context of a much wider study, that participants with high 'Need for Cognition', a measure of enjoyment of and reliance on conscious deliberation, pass significantly more utilitarian judgments on moral dilemmas than the rest of the sample. This perfectly fits with Greene's view.

There are two other results, based on manipulation of the judgment task, that support Greene's view. Valdesolo and De Steno (2006) showed to participants a clip from a funny TV program[34] and reported that participants were significantly more likely to approve of pushing the bystander down to her death in *Footbridge* after experiencing hilarity than in the control condition (no clip). Greene interprets mirth as an emotion countering the negative affect generated by System1 at the idea of committing an instance of personal violence. Hence, more mirth would mean less negative affect and more room for System2 to drive the final moral judgment concerning *Footbridge*.

Suter and Hertwig (2011) have recently either put participants under time pressure (they were given a fixed amount of time to read the dilemma and pass a verdict on a Likert scale) or asked them to be make judgments about a given scenario as soon as possible. In both cases people were more likely to make non-utilitarian decisions for high-conflict dilemmas under time pressure than in control situations where no time pressure was exerted. This dovetails with Greene's idea that System2 is cognitively expensive and slow. Time pressure can easily count as a cognitive load and hence a prevalence of System1 judgments under time pressure conditions is predicted by Greene's model.

Then there is the evidence from Greene and his lab. Greene and co-workers carried out four important experiments: although the fMRI experiments (2001, 2004) are the ones that made Joshua Greene "rich and famous," the behavioral experiments he conducted later (2008, 2009) are much more important for the dual-process model than the fMRI studies[35].

[34] "Saturday Night Live"

[35] This is Greene's opinion too. "My fMRI research was designed to test a psychological theory, and that theory is in no way bound to the technology that was first used to test it." (2010, 6). In personal communications he repeatedly stressed this point.

Greene et al. (2008) is a cognitive load experiment. Greene analyzed just high-conflict moral dilemmas, i.e. dilemmas that involve personal force and in which personal force brings about positive consequences from the point of view of aggregate welfare. The cognitive load consisted in a stream of digits crossing the screen: participants were asked to press a key each time they saw the digit 5 and they had to do that while they were reading the scenarios texts and passing judgments. Therefore, the experiment had two conditions: under load and in absence of load. It was predicted that (1) participants under load would make less utilitarian judgments than in the absence of load; (2) participants would take longer time to make utilitarian judgments under load than in absence of load, since cognitive load depletes cognitive resources needed to System2, while not stopping System1 from working; (3) even in absence of load the Response Times (RT) for utilitarian judgments in high-conflict dilemmas should be higher than those for non-utilitarian judgments. Just one of these three predictions is supported by the experiment. Greene and co-workers found that utilitarian judgments under load are indeed slower than in absence of load in a selective way (i.e. this effect does not extend to non-utilitarian judgments)[36]. This means that cognitive load slows utilitarian judgments *only* and it does nothing to deontological judgments. This seems to show that utilitarian judgments are linked to (narrowly) cognitive mechanisms, whereas deontological judgments are independent from them. This in turn buttresses the dual-process model of moral cognition. However, Greene and his co-workers found no effect of load on the number of utilitarian judgments made by experimental participants, nor any effect of judgment on RT in absence of load in the full sample. In other words, subjects under load do *not* make *less* utilitarian judgments than subjects in the control condition and utilitarian judgments are in general *not* slower than deontological judgments. A difference in speed would be expected if utilitarian judgments stemmed from reflection as posited by the dual-process model. These two null results are quite strong counter-evidence for the dual-process model, as Roskies and Sinnott-

[36] In statistical terms, Greene et al. (2008) found a significant interaction between judgment and RT in the two conditions.

Armstrong (2008) and then Berker (2009) remarked[37]. Nonetheless, the RT interaction between judgment and cognitive load stands as evidence in favor of the model, since it is a specific result. The RT difference in absence of load is an important prediction of the dual-process model, since utilitarian judgments in absence of load are thought to be the outcome of System2 overriding System1 and this necessarily takes more time than System1 acting directly to yield a deontological judgment. Greene et al. (2001) published RT data to the effect of such a RT difference but Greene later admitted (Greene 2009; Greene et al. 2008) that the RT results in the 2001 paper were due to very fast non-utilitarian responses to some scenarios in which the consequentialist response was completely unpalatable[38] and that thus do not properly count as 'dilemmas'.

The 2009 experiment tries to identify what drives System1. The initial hypothesis, which can be found in Greene et al. (2001) and in Haidt and Greene (2002), was the so-called ME HURT YOU paradigm: System1 would react to an action if (1) it is likely to cause serious bodily harm; (2) to a particular person; (3) the harm does not result from the deflection of an existing threat onto a different party. However, there are some actions, as Greene (2008b) recognizes, that respect these conditions but not trigger the fast system, such as Kamm's Giant Lazy Susan scenario[39] (Kamm 1996, 154). Hence, the ME HURT YOU paradigm was abandoned and the 2009 study was run to find an answer to the factors problem: what factors of a situation does System1 respond to? The answer is the following:

[37] Sinnott-Armstrong (personal communication) still thinks that the 2008 null result shows that Greene's overall theory just tells a part of the story about the machinery for moral judgments.

[38] The most notorious case is the *Hired Rapist* dilemma, in which we are told of a husband whose wife has become estranged. In order to regain her affection, the husband can hire a rapist to rape her. Then, after hearing the horrible news, the husband would return swiftly to her side, to take care of her, and comfort her. She will thus once again appreciate him. It is obvious that almost no one decides to hire the rapist. The problem with the RT was pointed out by McGuire et al. (2009) and by Berker himself (2009), but had already been noticed by Greene's co-workers before the publication of McGuire's and Berker's articles. Moore et al. (2008) tried to replicate the 2001 RT data (even though with different, better dilemmas) and failed to do so.

[39] In this scenario case, a runaway trolley is heading toward five innocent people who are seated on a giant lazy Susan. The only way to save the five people is to push the lazy Susan so that it swings the five out of the way; however, doing so will cause the lazy Susan to ram into an innocent bystander, killing him. Lab experiments briefly reported in Greene (2008b, 108) showed that participants tend to respond in a consequentialist way to the giant lazy Susan variant of the Trolley dilemma.

intention to harm and personal force[40]. The 2009 article comprised more than 15 dilemmas, featured a big sample size, and showed that both intention and personal force have an influence on judgments. Neither of them is a necessary or a sufficient condition for eliciting emotional condemnation. In some cases intention to harm alone is sufficient to activate System1, in others personal force is sufficient, under still others circumstances both are required. Hence Greene's and co-workers' final hypothesis is that System1 "operates over an integrated representation of goals and personal force - representations such as 'goal-within-the-reach-of-muscle-force'." (Greene et al. 2009, 370). The experiments reported in this article count as evidence for the dual-process model because they extend previous results to many different dilemmas and to a big sample, pin-pointing at the same time the factors that elicit the activation of System1.

Moving to the fMRI experiments, Greene et al. (2001) is admittedly a weak paper. The dilemmas are too heterogeneous to yield credible results. The fact that most criticism attacked this paper is not casual. It was a bold attempt, managed to get published in one of the highest impact factor scientific journals in the world, and opened the field to neuroimaging investigation. But no matter how seminal it was, relatively little can be saved of it, apart the fact that reading 'personal' moral dilemmas (following the ME HURT YOU paradigm) causes the BOLD activation of 'emotional' areas of the brain, which is far from surprising since the actions described are often disturbing. Starting from the 2004 article, Greene focused on high conflict dilemmas. The key result of the 2004 paper is that utilitarian responses to difficult or high-conflict moral dilemmas[41] bring about a hemodynamic activation in areas related to cognitive control, i.e. the DLPFC 'accounting department' and the Anterior Cingulate Cortex (ACC). However, both in the 2001 and in the 2004 papers there are strange BOLD activations that are not easy to explain away, e.g.

[40] Nonetheless, there is an important empirical result that qualifies this answer. Nichols and Mallon (2006) show that the different reactions to *Switch* and *Footbridge* are conserved when the entities damaged by the agent are not human beings, but china cups. This means that personal force and intention are important even when they do not constitute counts of battery or violence against sentient beings.

[41] Those which pit aggregate welfare and strong emotions one against the other.

DLPFC firing up less in personal dilemmas than in impersonal dilemmas and the PCC, an alleged 'emotional' area, firing up when utilitarian responses to personal dilemmas are made. Summing up, the fMRI evidence that made Greene hit the headlines is not particularly important to the survival of the dual-process model as a working hypothesis in experimental moral psychology.

The relevance of the fMRI results by Greene and colleagues is further reduced by fMRI itself being a correlative technique. It is not good at establishing causal connections, because from BOLD signaling alone it is not possible to say whether the relationship between the activated regions and a mind function is of necessary condition, sufficient condition, or a mere correlation. In a standard (i.e. non multi-voxel pattern analysis[42]) fMRI experiment, subjects are shown a stimulus and a control. They are normally similar, except for one item that differs and that, according to the researcher's theory, coincides with the mind function under examination. After data acquisition the researcher subtracts the activation at the moment of the control (e.g. neutral face) from the activation at the moment of the stimulus (e.g. scared face). Then regions that show a statistically significant activation after the subtraction are looked for. It is then possible to make an inference of this kind:

Premise 1: condition C1 (scared face) correlates with a BOLD activation in region R1 (amygdala);

Premise 2: function F1 (vision of fear in other human beings) correlates to condition C1 (scared face);

Conclusion: function F1 (vision of fear in other human beings) correlates with a BOLD activation in R1 (amygdala).

This is the so-called forward inference, or inference from function to structure. It is generally considered as legitimate. But there is also another kind of inference that is common in neuroscience: reverse inference, or inference from structure to function.

[42] For multi-voxel pattern analysis, cf. Haynes and Rees (2006).

Premise 1: function F2 correlated in previous experiments with an activation in R2;

Premise 2: condition C2 correlates in this experiment with an activation in R2;

Conclusion: condition C2 correlates in this experiment with function F2.

For example, we might derive an engagement of the function 'experiencing fear' from an amygdalar activation even if the experimental condition does not yield particular cues to think that fear is active in that case. However, the conclusion of a reverse inference is not sure. The inference I used as an example is indeed very dubious. The validity of a reverse inference depends on the specificity of the activation of R2. If R2 is also activated by functions F3, F4 and so on, the conclusion is invalid. It is quite difficult to build specific correspondences between functions and regions, but this becomes easier when little areas of the brain are taken into account. Hence, this problem would be eased by increasing the strength of the magnetic field used in MRI. This would allow for a smaller voxel size while keeping the signal to noise ratio constant, so that smaller areas of the brain could be highlighted. But an increased number of Tesla also creates artifacts and problems such as vertigo (Theyson et al. 2007), so that the issue is unlikely to be solved through technical innovations only. As a result, the legitimacy of reverse inference in neuroscience is contested.

Poldrack (2006; Poldrack and Wagner 2004) argues that reverse inference should be considered as a probabilistic inference and that it can be useful to create hypotheses that must be tested through forward inference in subsequent experiments. For example, if in a written word classification task a researcher finds an unexpected activation of a motor area, she can try to devise a theory to explain this activation and then test it through a different task and forward inference.

Henson (2005, 2006) tries to save reverse inference by claiming that current cognitive neuroscience uses a one-to-one function-to-structure correspondence as a working hypothesis. We should assume this hypothesis, use reverse inference *as though* it was justified and then judge the assumed working hypothesis on the basis of the results, i.e. of

the empirical success of the research program at explaining empirical phenomena. However, some experts (e.g. Pessoa 2008, 155; Price and Friston 2002, 2005) claim that one-to-one mapping in the brain is impossible because we have a trove of empirical findings showing that most areas of the brain are involved with multifarious mental functions. A further problem comes from the poor clarity in defining and systematizing psychological concepts. Many psychological concepts have lots of synonyms (e.g. the notorious triad "executive function", "working memory", "cognitive control") and we lack a list of functions we want to map onto the brain. If the aim of cognitive neuroscience is to create connections between brain structure and mental functioning, it seems that both sets must be carved at the joints, especially if we want to make experimental psychology incremental, i.e. if we want to build new scientific knowledge on the accumulation of single results from studies. A necessary condition if we are to do this is to make studies comparable and to build psychological ontologies, theories "about the structure of the mind that specify the component operations that comprise mental function" (Poldrack 2010; cf. also Price and Friston 2005). At any rate, it is exceedingly unlikely that a good ontology for mental processes, coupled with a good anatomical ontology for the brain, would justify the working hypothesis of a one-to-one mapping between functions and regions. However, Henson specifies that his proposed one-to-one mapping applies only to entities on the psychological or neuro-anatomical ontologies that lie on the same level of generality. Hence, there will be a one-to-one mapping between psychological functions and networks of brain regions at a certain level of detail. In other words, a general psychological function will map in a one-to-one way with a large, complicated network of neural structures, but at the same time a given general psychological function will map in a one-to-many way with smaller groups of neural regions that are more specific. So, it seems that the construction of precise ontologies both in the psychological and in the neural domain is essential to decide this debate.

Machery (forthcoming) distinguishes two interpretations of reverse inference. According to the Bayesian interpretation, a particular pattern of brain activation E supports the hypothesis that psychological process P is recruited by a given task T if and only if the probability of the occurrence of E if P is recruited is higher than the probability of E's occurrence if P is not recruited. For example, a bilateral amygdalar activation supports the hypothesis that the emotion 'fear' is recruited by the current experimental paradigm that involves judgments on gruesome moral scenarios if and only if the probability of the occurrence of a bilateral amygdalar activation if fear is recruited is higher than the probability of witnessing a bilateral amygdalar activation if fear is not recruited. Given that a single pattern of brain activation (e.g. bilateral amygdalar activation) is caused by many psychological processes, the probability of occurrence of E when P is *not* recruited is likely to be high, so that Bayesian reverse inferences are likely to be false. Machery contrasts the Bayesian interpretation with a 'likelihoodist' interpretation, according to which E provides evidence for the hypothesis that T recruits some psychological process P1 over the hypothesis that T recruits another psychological process P2 if and only if E is more likely to be found when P1 is recruited than when P2 is recruited. Machery backs this interpretation and notices that it is restrictive relative to current practice, since it solely works in comparative cases. As a matter of fact, current practice uses reverse inference in traditional experiments (as opposed to automated meta-analyses such as those proposed by Yarkoni et al, see below) to explain (albeit speculatively) BOLD activations that are not expected given the structure and content of the task. Since these instances of reverse inference are not necessarily comparative, most of them ought to be seen as unwarranted.

Klein (2011) sees reverse inference as an instance of inference to the best explanation (IBE) (Lipton 1991) and claims that it is an invalid form of IBE because regions are pluripotent, i.e. associated with different mental processes at the same time. Instead of reverse inference, Klein proposes cross-domain abduction, which consists in looking for a theory that explains the activation of a region across all (or at least as many as possible)

tasks in which the region is activated by invoking a minimum amount of functions. Instead of considering only one task, as standard reverse inference does, we should take account of all of the tasks in which the region is involved; then we should minimize the number of functions we invoke to explain the BOLD activations. Klein takes as an example of cross-domain abduction the treatment Price and Friston (2005) made of the activation of the Posterior Lateral Fusiform (PLF) gyrus. The PLF is activated by a long list of experimental paradigms: viewing words, picture naming, making unprimed semantic decisions, decoding Braille, and so on. By taking into account all the neuro-imaging data available, Price and Friston propose that PFL is mostly connected with sensorimotor integration, since these activities require such a function.

This proposal is not very far away from that by Yarkoni et al. (2011). In a seminal paper, Yarkoni and co-workers claimed to have built an automated classification software that can extract BOLD activation foci from hundreds of already published neuroimaging articles and draw automated reverse inferences on the basis of these data with pretty good accuracy. In other words, Yarkoni and colleagues (among whom Poldrack) have found an automated way to probabilistically infer the mental function corresponding to a given set of hemodynamic activations. For instance, the BrainSynth system allows to insert a psychological term, such as 'pain', into the system. Then the system automatically carries out a meta-analysis of published material, retrieves the stereotactic coordinates of the relevant activation foci, and finally yields a brain map that shows what is the possibility of engagement of the different Brodmann areas for the psychological concept provided. This corresponds to forward inference, i.e. we move from function to structure. On the contrary, it is possible to feed a BOLD activation map into BrainSynth and retrieve the probabilities of engagement of each psychological function given the map. Selection of the psychological term with the maximal probability provides a neat reverse inference, i.e. an inference from structure to psychological function. If this is confirmed and replicated, this automated system would not only pave the way to reliable reverse inference, as I have just

said, but also to reliable mind-reading, i.e. decoding of BOLD patterns to mental states in an open-ended way without training or previous knowledge of the "ground truth," i.e. on the basis of previous literature only. The system seems to be able to associate psychological terms to BOLD patterns simply drawing on already published foci and with a pretty good level of reliability. This would be a revolution in cognitive neuroscience.

Summing up on reverse inference, the proposal by Klein makes sense and the work by Yarkoni and co-workers suggests that prior literature can effectively be mined to get information about function-to-structure mapping. So there are in principle no problems in making reverse inferences, provided that these two conditions are respected: (1) their probabilistic nature must be taken into account; (2) previous literature must be accessed in a non-anecdotic, systematic way. Experimental moral psychology should not assume a bijection between structure and function, since the latter is very unlikely to stand the test of experimentation. To increase the specificity of the correlation between structure and function, experimental moral psychologists could use, in addition to fMRI meta-analyses, other techniques, such as repeated Transcranial Magnetic Stimulation (rTMS)[43] and lesion studies. These techniques allow to verify if a region of the brain is *necessary* to the performance of a mental function. fMRI alone is unable to do so. By combining lesion studies, TMS and fMRI on the one hand, and deploying the automated meta-analytic tools proposed by Yarkoni and colleagues on the other hand, real causal connections between brain regions and functions could be established. Furthermore, traditional fMRI studies should avail themselves of much bigger sample sizes, as most neuroimaging studies are underpowered (in terms of statistical power, cf. Yarkoni et al. 2010). It must be stressed that the probabilistic nature of reverse inference is not *per se* a problem – entire areas of empirical science are probabilistic and in particular experimental psychology is such, since it chiefly relies on Null Hypothesis Significance Testing. Even forward inference in fMRI are essentially statistical, since Student's T tests are necessary to establish which voxels are

[43] At some specific frequencies this kind of magnetic stimulation can simulate brain lesions, as it can effectively disrupt neural functioning in a transient and localized way.

significantly activated in a given fMRI contrast. Caution must be addressed only to the idea of a rigid one-to-one mapping between mental functions and regions that comprise millions of neurons. The fMRI experiments by Greene heavily rely on reverse inference, as they classify regions into 'emotional areas' and 'rational areas'. If reverse inference turned out to be illegitimate, those results would have little import for experimental moral psychology. However, the dual-process model is backed by much more empirical evidence than those two papers, so that the destiny of Greene et al. (2001, 2004) is largely orthogonal to the destiny of the dual-process model of moral cognition.

Going back to the work by Greene and colleagues, there are finally three experiments that are worth mentioning.

First, Cushman et al. (2012) showed that Total Peripheral Resistance (henceforth TPR), a measure of physiological arousal that is a function of both cardiac output and mean arterial pressure, is a significant predictor of responses to moral dilemmas. Participants with higher TPR are significantly more likely than the rest of the sample to condemn the action of throwing a man out a lifeboat in order to save the other passengers from drowning. Hence, more emotional arousal leads to more deontological judgments.

Second, in Paxton, Ungar, and Greene (2012) experimental participants had to perform the Cognitive Reflection Test (henceforth CRT). The CRT consists of three questions that elicit incorrect, intuitive responses, which can be overridden by correct, reflective responses through the application of basic math, e.g.:

"A bat and a ball cost $1.10.

The bat costs one dollar more than the ball.

How much does the ball cost?"

Nearly everyone has the intuition that the ball will cost $0.10. However, the solution of a linear equation and a bit of reflection are sufficient to discover that the correct answer in $0.05. Subjects that can answer correctly to at least one of the CRT items are considered more reflective than the others. Paxton and co-workers showed that reflective participants

as defined by the CRT are more likely to give utilitarian responses to high-conflict moral dilemmas, thus buttressing the idea that utilitarian judgments are linked to controlled, reflective cognitive processes.

Thirdly, Amit and Greene (2012) divided participants into visual and verbal sub-groups through a visual-verbal working memory task. Participants that were better at associating pictures were labeled as 'visual', whereas the ones more apt at categorizing words were labeled as 'verbal'. The experimental hypothesis that drove Amit's and Greene's experiment is that visual imagery should be correlated with the visualization of means in high-conflict dilemmas. When one imagines the *Footbridge* case, one mentally visualizes the man that gets thrown down the bridge and not the five human lives that are thereby saved. Furthermore, visual representations are more emotionally salient than verbal representations. Hence, Amit and Greene hypothesized that people who visualize more are more likely to engage emotions, to concentrate on the rights of the bystander in *Footbridge*, and to pass deontological judgments. The result of the experiment confirmed these hypotheses: visual participants are more 'deontological'. This nicely dovetails with the result by Caruso and Gino (2011), who showed that keeping one's eyes closed increases the degree of mental simulation one carries out and renders moral judgments more emotional and more extreme. Hence, there is an overall correlation between mental imagery, emotional arousal, and harshness of moral condemnation towards rights violations[44]. To further strengthen their finding, Amit and Greene assigned to the participants two different cognitive load tasks, one visual and another verbal, to be executed during the rating of moral scenarios. The visual load was expected to tax the capacity of performing mental imagery, thereby reducing deontological judgments. The verbal load was expected to work like the cognitive load used by Greene et al. (2008). The latter load had (quite curiously) no effect, whereas the former actually made judgments

[44] Notice that Caruso and Gino (2011) did not use high-conflict moral dilemmas, but simple descriptions of violations.

more utilitarian relative to no interference, thereby confirming Amit's and Greene's research hypothesis.

To summarize this overview of experimental results in favor of Greene's dual-process view, it is admittedly a wealthy amount of empirical evidence. The fMRI evidence, that has been the focus of so much controversy because of both the relative novelty of the technique[45] and the general "neuro-hype" that characterizes the later years, is not at all essential to the model or even particularly important. The dual-process model is a thesis in experimental moral psychology and not primarily in neuroscience of morality. The body of evidence for the dual-process model has one chief limitation: it comes almost exclusively from "trolleyology." If "trolleyology" came up as an unsound or misconceived way of exploring the moral domain, the dual process model would be scientifically dead. It is not a case that Greene and co-workers advocate the usefulness of dilemmas as a tool in experimental moral psychology quite forcefully (Cushman and Greene 2011): their whole research program depends on the viability of this approach. However, there are many scholars who think that trolleyology is not a good way to investigate moral behavior from the experimental point of view. In the next section I examine, among others, some of these critiques.

2.3. Critiques of Greene's dual-process model

Greene's dual process view has so far been the topic of much criticism. I discuss some of it here and some in the next chapter, where I deal with alternative views to Greene's proposal.

As I have written above, one important criticism tries to show that moral dilemmas are not a viable method in experimental moral psychology. The criticism is often made by neuroscientists and philosophers close to virtue ethics (Casebeer 2003; Churchland 2011). These researchers often follow a reading of history of moral philosophy by McIntyre (e.g.

[45] The BOLD contrast was invented in the early 1990s by S. Ogawa (cf. Ogawa et al. 1990).

1982), according to which the Enlightenment movement reduced the thick (i.e. teleological and Aristotelian) concept of ethics to a thin concept in which attention was moved from the agent to the action and contexts were virtually obliterated from consideration. According to an Aristotelian conception of ethics, the character of the agent is central and a virtuous agent is able to deftly respond to diverse demands stemming from the precise context in which she is embedded. Aristotelians therefore evaluate this shift in a negative way and do not look kindly upon attempts at creating very artificial situations like those that appear in Trolley-like cases. They argue that artificial cases tell us little about what morality is in everyday life and yield a warped view of human moral behavior.

Churchland (2011, 110) notices that some moral judgments vary a lot if the context is changed: drinking "fresh apple juice out of brand-new hospital bedpan" feels disgusting in the lab, but it is not disgusting when you are dying of thirst in a desert. However, this criticism is not hard to defuse. Churchland could surely grant that in most contexts (e.g. while one is parking her car, when one is watching a football match, while one is cooking, etc.) the idea of drinking apple juice from a brand new bedpan is perceived as disgusting in Western cultures. Thirst would be an experimental confound here, an additional factor that significantly changes the psychological process we want to investigate, i.e. the one that takes place in most contexts. The same holds for the case put forth by Casebeer (2003, 646) that stealing weapons from terrorists is morally good. Casebeer wants to underline the context-dependency of moral judgments in this way: stealing is morally bad in most circumstances, but the context can exert such a powerful effect on normative standards that under specific circumstances this action becomes acceptable. However, war and police operations are precisely contexts in which social groups substitute their usual moral categories with others: the morality of a soldier (bravery, honor, obedience) is not the morality we use while we are shopping at the super-market and the former does not say much about the latter. Again, here the context of a police operation against a terrorist group would change the psychological process the experimenters want to untangle. What we

think when we evaluate the action "stealing a lollipop from a child" is not the same as our mental process when we evaluate the action "stealing explosives from terrorists," even if one is not a police officer. Casebeer also maintains that moral cognition is emotionally loaded and social, but this does not necessarily count as an objection against hypothetical scenarios. Affective and social psychology are burgeoning fields of empirical investigation that study emotions and social bonding inside the clean walls of the lab, i.e. leaving aside the real-world context. These lines of research often create hypothetical cases that seldom happen in real life (cf. for example the famous – or notorious – experiments by Milgram 1963). Of course ecological validity[46] ought to be pursued as much as it is feasible to do so, but it cannot constitute an in principle objection neither to empirical inquiry, nor to some specific experimental strategy, unless more specific claims against distinct experiments are made. I concede that social and affective neuroimaging are more challenging than social and affective experimental psychology in general, but even in this case ecological validity is in general preserved to such an extent that experiments are considered valid by the relevant scientific community. It is difficult to simulate social bonds and emotions inside a 3T scanner that makes a lot of noise and in which only one person at the time can stay, but it can be done by careful selection of the stimuli and eventually with the costly technique of hyper-scanning, i.e. using two scanners, two participants at the time, and allowing the latter to communicate to each other while both of their brains are scanned (Montague et al. 2002)[47]. The most interesting criticism made by Casebeer is that moral knowledge[48] is a knowledge-how and not a knowledge-that, and the paradigms used by Greene (and by many other experimental moral psychologists, I have to say) deal with moral knowledge as it was knowledge-that. Moral knowledge would be

[46] Ecological validity is the homogeneity of the experimental environment (stimuli, context, etc) relative to the parts of the real world it wants to model. For example, if you want to uncover what happens in a participant's brain when she conceals info about a crime she committed in the recent past, it is preferable to make her commit a mock crime (e.g. stealing something) or to give her the possibility to be dishonest. In this way ecological validity will be high, or at least higher than using other paradigms. See for instance Kozel et al. (2009).

[47] Notice however that hyper-scanning is rare. On a purely anecdotic level, the only paper I have ever read which uses this technique is Krueger et al. (2007).

[48] Assuming *arguendo* that this disputed concept makes sense.

knowledge-how because moral skills enable good navigation in a complex social environment where the individual is a moral agent as well as a moral judge. She decides at the same time what judgments to make and what actions to carry out. It might be the case that moral knowledge is actually knowledge-how. In this case we should change our experimental paradigms. However, some experimenters have already asked participants to carry out physical simulations of morally relevant actions in the lab (cf. Cushman et al. 2012), for instance pretending to hit a person. In this specific case, participants were clearly informed that they were hitting a fake hand or using a fake knife, but any way subjects were asked to act, not to judge only. Of course the strict limitations to human experimentation stop researchers from going too far on this kind of experiments, and rightly so. Even though the possibility that moral knowledge is knowledge-how should not be neglected, I do not know at present of the existence of any evidence indicating that this possibility is actual.

A third criticism of this kind is leveled by German psychologist Gerd Gigerenzer (2010). He is skeptical about moral dilemmas because he criticizes RCT by drawing on the works by Nobel-laureate Herbert Simon (e.g. 1955, 1956) on bounded rationality. In his opinion RCT and the associated normative ethical theory of maximizing consequentialism can work only in "small worlds," realities in which all the probabilities and the values of outcomes are known to the agent. This knowledge is possible exactly because these settings are artificial and lack complexity. The real world in which humans live is not a small world, though. It is exceedingly complex and humans living in it systematically lack information, so that they have developed decision making strategies that are conducive to survival and reproduction even in absence of significant pieces of information. Bounded rationality embodies these decision making strategies. Hence, in Gigerenzer's opinion experiments that are carried out in the small world of the lab would tell little about the large world outside its walls. Put in this extreme form, Gigerenzer's argument seems to doom the whole of experimental psychology, including his own work, since most of

psychology is done in the lab to allow for precise elimination of confounding variables. If a more charitable reading is adopted, this claim seems to be reducible to an appeal to ecological validity, an appeal that is well taken but that is insufficient to condemn experimental strategies or paradigms unless more specific, circumstantiated claims are made.

A fourth criticism is leveled by Jana Schaich Borg and colleagues (2006) who investigated, through fMRI and among many other moral phenomena, the neural correlates of the Doctrine of Doing and Allowing (henceforth DDA)[49], a typical deontological tenet utilitarians regularly reject. The regions involved with the DDA were markedly cognitive areas such as the DLPFC. This is evidence against the idea that deontology is strongly connected with System1, assuming *arguendo* that a thing such as System1 exists. Schaich Borg and co-workers suggest that some deontological responses are mediated by reason, whereas other deontological responses are mediated by emotion: no specific mapping between reason/emotion and deontology obtains.

Some different but however important criticisms of Greene's descriptive work have been made by Oxford philosopher Guy Kahane, arguably the most perceptive of Greene's critics. In Kahane and Shackel (2010) three criticisms are leveled.

First, there is a criticism on words. Kahane and Shackel argue that Greene and colleagues have asked participants whether hypothetical actions are "appropriate" or "inappropriate," but this does not allow the researchers to understand what kind of moral concepts the participants are deploying. Are participants that answer "appropriate" thinking that the action is mandatory? Are they thinking that the action is merely permissible? Are they thinking that the action is praiseworthy but not mandatory? The argument is not new, as you can find it in a similar form in Kamm (2009). This point is about the wording of the question that is addressed to the participant, not about the wording of the scenario. This criticism has been defused from the empirical point of view by O'Hara, Sinnott-

[49] The DDA is the idea that one is more justified at inflicting harm through an omission than through an action having the same effect as the omission. For more information, cf. Quinn (1989a).

Armstrong, and Sinnott-Armstrong (2010), who showed that studies whose questions featured different wording can be legitimately compared because subjects respond roughly in the same way when different (but similar) questions about the same scenario are made. Participants do not perceive as different concepts which professional moral philosophers deem different. There is another wording problem, though, that Kahane and Shackel do not mention. It has to do with the wording of scenarios. This wording can have serious effects, as famously noticed by the late Amos Tversky and Nobel-laureate Daniel Kahneman (1981). Describing the effects of a public health policy in terms of either deaths or lives saved from death itself[50] dramatically changed participants' responses. This type of wording effect was already discussed by the early experiments on Trolley-like scenarios carried out by Petrinovich, O'Neill, and Jorgensen (1993, 476) and by Petrinovich and O'Neill (1996, 152) and the result was that this wording effect is a potent explanatory factor for participants' decisions, as it can account for roughly 25% of variance in responses. Greene's scenarios do not seem to show this specific problem, but any way wording of hypothetical scenarios must be controlled in experimental moral psychology. Concluding, both criticism on wording do not seem to hit Greene's experimental program that hard.

Secondly, Kahane and Shackel criticize the choice of the dilemmas. This criticism is widespread (Berker 2009; Kamm 2009; Mikhail 2011; Moore et al. 2008). There are two different criticisms, though. The former is about the 2001 dilemmas (starting from the 2004 paper Greene has analyzed just a subclass of personal moral dilemmas). As Moore et al. (2008) correctly notice, the personal and impersonal dilemmas used in the 2001 paper involved different kinds of violations (killing v. stealing, for instance), different levels of probabilistic reasoning, different levels of difficulty since some scenarios were not dilemmatic at all, and different lengths in terms of number of words. Moore et al. managed

[50] The notorious 'Asian flu' case.

to replicate Greene's personal/impersonal effect[51] using refined dilemmas, even though they failed to spot a link between a measure of working memory and utilitarian judgments. This null result counts as a piece of evidence against the dual-process model. That being said, Greene partially addressed this problem both by restricting his set of dilemmas to the high-conflict list first used by Koenigs et al. (2007) in his VMPFC experiment and by using completely different dilemmas from the 2001 ones in subsequent works (i.e. Greene et al. 2009). To my knowledge, changing the dilemmas has never overturned one of Greene's results, except the 2001 RT result. At any rate, the 2001 dilemmas are technically very bad and should be jettisoned and forgotten. It is not difficult to come up with more refined, better crafted scenarios (for instance Cushman et al. 2011; Moore et al. 2008). Furthermore, a second criticism addresses the difficulty of using the dilemmas to map a distinction between utilitarian and deontological responses. It is the same difficulty discussed above while commenting on Koenigs et al. (2007) and, though reasonable, it is empirically defused by the re-analysis carried out by Koenigs and his co-workers while responding to Kahane and Shackel (2008).

The third and last criticism by Kahane and Shackel (2010) is the one about moral theories. Patterns of responses to dilemmas, in their opinion, cannot be used to attribute to participants the belief in moral theories. In the case of utilitarianism, a person can be considered a maximizing act utilitarian if and only if she believes that the *only* thing that determines whether an act is morally right is whether it maximizes aggregate well-being. Experiments by Greene and others are insufficient to prove that any of the participants endorse such a belief. This is philosophically correct. It seems that this *is* what we want to know when we want to ascertain whether someone is a card-carrying utilitarian. But is Greene interested in the card-carrying utilitarian? Or is experimental moral psychology interested in who is a card-carrying utilitarian? Not necessarily. The focus of Greene's model is on judgments, i.e. the output of the two alleged systems. As far as the analysis

[51] In other words, participants make more utilitarian judgments responding to impersonal rather than to personal moral dilemmas.

stays at the level of judgments, we do not need to be concerned with the attribution of endorsement of substantive ethical theories, whose work is justifying moral judgments. But at the same time Greene's model contains a claim on the *theories* being psychological natural kinds. It seems impossible to back this claim from the empirical point of view unless *experiments on theories* are made and Kahane's criterion for attribution of belief in a substantive ethical theory is used. But there are little or no experiments on the issue of moral justification in experimental moral psychology, with the possible exception of Cushman, Young and Hauser (2006) and Hauser et al. (2007). Studies on moral justifications are so rare because they need to rely on self-reports, and self-reports have been looked on with suspicion in social psychology since the famous review by Nisbett and Wilson (1977) that shows that they are normally unreliable. In absence of solid experimentation on moral justifications little can be said about normative ethical theories from the experimental point of view. So, when Greene writes that "making 'characteristically consequentialist' and 'characteristically deontological' judgments requires no explicit or implicit commitment to consequentialist or deontological theories" (forthcoming, § I), he makes a reasonable point, but he does not realize that by conceding this point he is emptying his claim about substantive theories as psychological natural kinds of its alleged empirical backing. In this way he secures the part of his work that investigates judgments, but *de facto* jettisons claims about theories without realizing it. The problem gets even more serious if we understand, as Kahane and Shackel correctly suggest[52] together with Frances Kamm, that it is very well possible for several non-consequentialist theories to justify the judgment that it is morally permissible to push the bystander down the bridge to her death in *Footbridge.* It is for example possible for a card-carrying non-consequentialist to consider the right to life of the five more weighty than considerations against voluntary and personal violence. Deontology has been admitting

[52] Kahane and Shackel (2010, 575) are in my opinion correct when they argue that "There is, after all, a considerable overlap between what utilitarianism and many deontological theories require or forbid, given that nearly all deontological theories recognize the moral significance of outcomes, and duties of beneficence will very frequently prescribe the same acts prescribed by utilitarianism."

conflicts between *prima facie* rights and resolution thereof at least since the time of Ross (2002/1930), so that the possibility for non-consequentialism to justify the 'characteristically utilitarian judgment' in *Footbridge* should not come as a surprise. Concluding, it is useful to draw a distinction between a weak dual-process model, that simply deals with a System1 and a System2, and a strong dual-process model, that associates System1 to deontology, System2 to utilitarianism and makes the far-fetched 'natural kinds' claim (Berker, personal communication[53]). Greene seems to have some evidence for the weak version, but little evidence for the latter, strong version. To stress the same point, I also endorse the criticism by Dean (2010), according to which even assuming *arguendo* that Greene can attribute to deontology the emotional reactions to personal dilemmas, this would just represent a tiny part of a deontological ethical theory. As Dean notices, just one of the Ten Commandments is about the intentional use of personal force. The rest of that deontological theory has little to do with the experiments we are considering here, so that it is hard to infer any conclusions on the whole of that substantive ethical theory on the basis of these data.

There are another pair of objections that are worth mentioning. The first is made by Leben (2011) and is equivalent to the one by Fine (2006) against Haidt (2001) I discuss in Chapter 3. The rather Aristotelian point is that moral education and training can render automatic responses that at first required reflection and the deployment of reasoning. Habit and learning can create a 'second nature', i.e. automatic patterns of behavior that end up being almost as automatic as patterns that are regularly develop in ontogenesis, in a way that is robust to environmental perturbations. In order to apply the idea to Greene's descriptive model, one could say that the products of System2 can become System1-like responses under certain conditions, that have to do with repetition and reinforcement. For instance, a card-carrying utilitarian could decide to regularly override his aversion to personal violence and act as though personal and impersonal ways of bringing about harm

[53] This personal communication is a lengthy email by Berker addressed to Joshua Greene in response to Greene's (2010). Berker forwarded this email to me.

were morally equivalent. If this objection is correct, the inhibition of System1 and the systematic passing of judgments that bypass the personal / impersonal distinction would become automatic and unreflective over time. Conceding such a point would weaken the dual-process model because it would make the two systems more enmeshed and less distinguishable. How damaging to the model this objection is depends on how often this learning takes place, though. In other words, in order to understand whether the dual-process model is still descriptively useful after learning is taken into account, one needs to know how relevant it is in the overall economy of moral cognition. I leave this empirical question aside precisely because it is an interesting experimental issue on which, to my knowledge, no specific work has been done, also because it would require lengthy longitudinal studies. The objection seems to be plausible, but only empirical work can establish whether it actually holds and what scope it has.

The last objection comes from an experimental paper by Kahane and co-workers (2012). The question that drives this article is the following: are there deontological judgments that look very counter-intuitive, in the sense that just a few people make them? Kahane et al. argue that the dual-process model seems to be incompatible with such judgments, if they do exist. Deontological judgments should be intuitive in Greene's model. But there are some deontological positions that actually are very counter-intuitive in this sense. In the Anglo-Saxon world, deontological theories are regularly associated with the name of 18th Prussian philosopher Immanuel Kant. Kant is notorious for a short text in which he construes a complete prohibition of lying, even when this is required to protect some important values, such as a human life, or the well-fare of many people (Kant 1966/1797). That is to say that some forms of deontology prohibit white lies. Kahane has speculated that these forms of counter-intuitive, deontological judgments are driven by System2 and not by System1. He created a set of dilemmas that are labeled as Utilitarian Intuitive (UI), in which the characteristically utilitarian response is far more common than the deontological response. Among these dilemmas White Lie cases, that pit telling the truth

against significant psychological harm, are prominent. Then Kahane and colleagues showed these dilemmas plus Deontological Intuitive dilemmas from Greene's original set to experimental participants in a MRI scanner. Their results allegedly show that responses elicit no stable BOLD pattern according to whether they are utilitarian or deontological (i.e. according to content), but according to whether they are intuitive (i.e. partaken by most people in their social group) or counter-intuitive (i.e. relatively rare in the social group). On the basis of this result, Kahane and coworkers claim that the dual-process model is misconceived because the two alleged systems that would constitute the machinery for moral judgment lack identifiable neural correlates. Leaving aside technical issues with the interpretation of Kahane et al's fMRI results, which are fairly complex, there are at least two broader problems with this article. First, when Greene says "intuitive", he does not refer to agreement in a group. He refers to a cognitive process that has specific features and can be deployed under some circumstances and not deployed under some other circumstances. There are several dilemmas, such as *Switch*, in which the System1 (intuitive) response is a minority option from the empirical point of view, i.e. there are less participants that take the intuitive decision than the counter-intuitive decision. In *Switch* the cognitive process "intuition" is unlikely to be deployed due to specific features of the scenario. Kahane and coworkers classified responses as 'intuitive' or 'counter-intuitive' on the basis of the ratings of 18 independent judges, i.e. they used response ratings from a sample as classification criterion. So there is a definitional disagreement between Kahane et al (2012) and Greene on the concept of 'intuition'. In Kahane's opinion intuitive judgments are simply those that are made by the majority of the sample, whereas Greene makes use of Haidt's concept and sees 'intuition' as a fast, unreflective cognitive process that is sometimes coupled with emotional phenomena. To a certain extent Kahane and Greene are therefore talking past each other and the experiment by Kahane et al. (2012) appears to be rather orthogonal relative to the dual process model. Indeed, the latter allows for very widespread System2 (counter-intuitive in Greene's sense)

judgments and of very rare System1 (intuitive in Greene's sense) judgments. For example, in a version of *Footbridge* in which the consequence of sparing the bystander's life is the death of 500,000 humans due to the explosion of a nuclear weapon, the System2 (counter-intuitive in Greene's sense) response is very common in the sample. Secondly, in an article Joseph Paxton, Joshua Greene, and me are writing while I am typing this (September 8th 2012), we report a behavioral experiment in which we used the CRT on a White Lie case. According to Kahane's hypothesis, more reflection should induce people to become more deontological in that case, since the hardcore Kantian response to that dilemma is seen as mediated and counter-intuitive. On the contrary, if Greene's model is correct on this point, correct answers to the CRT (which indicate more reflection) should correlate with more utilitarian responses to a White Lie dilemma. It is true *by definition* that priming through the CRT reduces 'intuitive' processing in a Haidtian sense as described in Ch.1, since those judgments are (again, by definition) unreflective judgments. Our results confirm Greene's view, even though the effect is not very strong, possibly due to a ceiling effect: most of our participants already endorsed the utilitarian solution in absence of CRT and it was therefore difficult to make them *even more* "utilitarian."[54] However, when people reflect more, they tend to tell lies to avoid psychological harm to others. They do not become crusaders of truth for the truth's sake. This shows that White Lie cases are not a counter-example to Greene's double mapping between System1 and characteristically deontological judgments on the one hand, System2 and characteristically utilitarian judgments on the other hand.

Concluding the chapter, there is significant evidence for a weak version of the dual-process model, as defined above. The main point on which Greene's project fails is the thesis that utilitarianism and deontology are psychological natural kinds. Much additional evidence, a significant part of which ought to be cross-cultural, must be marshaled by Greene in order

[54] Given what I have written above about theories and the dual-process model, I use this term just in a metaphorical way for clarity's sake.

to make that very strong claim credible. Unless this empirical evidence is provided, Greene's 'psychological natural kinds' claim strikes as false.

In the next chapter I review theories in moral psychology that are partially or completely alternative to the dual-process model, in order to show that the dual-process view is not the only cock in the hen-house of experimental moral psychology.

Chapter 3: Alternative views in experimental moral psychology and their ethical consequences

In this chapter I review some descriptive theories in experimental moral psychology that are alternative to Greene's dual-process model. I also examine and discuss the normative consequences that their authors draw from them, if any.

Two concepts must be explained at the outset.

The first is the idea of a neuromoral theory. A neuromoral theory claims that knowledge of the machinery for moral judgments can help individuals make better moral judgments. This is the common ground of all neuromoral theories[55]. These theories diverge on two main issues:

(1) What does it mean to say that the human ability to make good moral judgments can be *improved*?

(2) How exactly can the scientific understanding of the machinery for moral judgments have a positive impact on this ability?

(1) has to do with the standard according to which moral judgments are good or bad, better or worse than others. There are several possibilities. A cognitivist may think that moral judgments are belief-like, doxastic mental states. If the cognitivist is also a realist, then these judgments are good or bad insofar as they track the posited moral facts, whatever they are. In this case, knowledge of the machinery for moral judgments could help us make moral judgments that are closer to the moral facts than the previous ones. If the cognitivist is not a realist, he may be an error-theorist. An error theorist is committed to the view that moral judgments are doxastic states that are universally false, since there are no moral

[55] There is a possible theory that is slightly different from the others and should be named "debunking neuro-normative" theory. According to this theory, the main normative implication of a correct scientific understanding of the machinery for moral judgments is that people *ought not* to make moral judgments. On this view, discoveries in psychology and neuroscience reveal that there is something fundamentally wrong with *all* moral judgments and that, as a consequence, moral thinking ought to be abandoned and ought possibly to be replaced by some other form of thinking. For the recommendation to be coherent, the 'ought' in "one ought not to make moral judgments" has obviously to be understood in non-moral terms.

facts. But the error theorist *can* say that some moral judgments are better than others. For instance, he can say that moral judgments, albeit universally false, are better if they promote stable and rewarding forms of cooperation between *H sapiens* specimens[56]. In this case, knowledge of the machinery for moral judgments could help us make moral judgments that lead to forms of cooperation that are more stable and efficient. Alternatively, the error theorist may uphold some other normative standard for moral judgments. A non-cognitivist thinks that moral judgments are non-doxastic mental states, that have little to do with beliefs. Also the non-cognitivist *can* say that some moral judgments are better than others. For instance, a good moral judgment may be a non-doxastic mental state that helps humans interact in ways that are likely to generate the greatest happiness for the greatest number, or in ways that enhance stable and fruitful forms of cooperation, and so on[57]. In what follows, I remain agnostic on the disputes among cognitivists of the realist strain, error theorists, and non-cognitivists.

What matters is simply to emphasize that different standards can be used to assess the relative goodness of moral judgments and that it is not *per se* evident which one is correct.

(2) has more to do with the structure of the machinery for moral judgments, so that the answer provided by each neuromoral theory to this question quite heavily depends on the description of the machinery it is linked to. Not all descriptive views in experimental moral psychology are linked to neuromoral theories, but some are. Neuromoral theories have recurring problems, and I highlight them on a case per case base.

The second concept that needs to be briefly addressed is the concept of moral domain, even though, as I have written in the Introduction, I do not delve into it. Descriptive theories in experimental moral psychology have many differences, but one important difference is that they draw the distinction between moral judgments / scenarios / problems / beliefs and their non-moral counterparts in different ways. In other words, different descriptive theories identify different moral domains. Hence, these theories are not exactly taking into

[56] This would be a prudential standard.
[57] This would be a prudential standard too.

account the same domain and are not trying to make sense of exactly the same set of experimental results. Some theories may describe a moral domain that is rather limited and narrow, i.e. containing a small amount of kinds of actions. Other theories may identify broader domains and "moralize" issues that other theorists do *not* "moralize." For instance, some theories, such as the one sketched by Turiel (1983), limit morality to the domains of inter-personal harm and fairness. Others, such as the tripartite view advocated by Shweder and colleagues (1997), is much more inclusive and features behaviors such as the quest for sexual purity, deference to established authority, and obligations due to high rank and status. However, some behaviors and related experiments are consistently perceived by theorists as part of the moral domain, at least to my knowledge. I know no moral psychologist that takes the deliberate killing of an non-consentient, healthy, adult human being as a non-moral issue. Truth be told, one theorist, Walter Sinnott-Armstrong, claims that the moral domain is irreducibly pluralistic, so that scientific investigation ought to examine phenomena that are much more fine-grained than generally understood "moral violations." I will briefly deal with Sinnott-Armstrong's theory (2008b, 2012) in § 3.6. Apart from Sinnott-Armstrong, there is some degree of coherence in the amount of empirical material that these psychological theories have to cope with. To conclude, the purpose of this chapter is twofold: on the one hand, to show that there are many other descriptive models than Greene's, and on the other hand, to discuss neuromoral theories that are connected to these alternative descriptive views.

3.1. Haidt: The Social Intuitionist Model and the Moral Foundations Theory

At the moment Jonathan Haidt is arguably the most influential experimental moral psychologist. His work has given rise to two popular books (Haidt 2006, 2012) and to several academic responses, mostly critical (Fine 2006; Paxton and Greene 2010; Pizarro and Bloom 2003; Saltzstein and Kasachkoff 2004; Suhler and Churchland 2011). I deal with some of these comments below. Haidt's work is composed of two theories that are

rather independent: the Social Intuitionist Model (henceforth SIM) and the Moral Foundations Theory (henceforth MFT). I will start with the SIM and deal with MFT later on in this section.

The SIM makes two main claims: (1) moral judgments are mostly intuitive, i.e. made without reflection, i.e. they are moral intuitions in the Haidtian sense I have discussed in Ch. 1; (2) moral judgments are a social practice in which the individual is continuously influenced by her fellow humans and her social milieu; in particular, moral reasoning regularly takes place at the social level (inter-personal discussion) and not at the individual level (solitary cost-benefit analysis before deciding how to judge a particular situation). The model was first expounded in Haidt (2001). It must be underlined from the outset that the SIM is a theory of moral judgment, and not (more broadly) a theory of moral decision making (Haidt and Bjorklund 2008b, 242). Moral judgment is a mental state in which a moral property of some sort is attributed to an action, an omission, or a character. It is typically in the Other perspective, i.e. a judge J evaluates actions, omissions, and character traits that are not carried out by J or that do not belong to J. However, there are also moral judgments in the Self perspective, although Haidt does not seem to take them into account. In contrast with this, moral decision making is much more complex, since it also includes selecting worthy goals, considering alternatives, and weighing different considerations (cf. Narvaez 2008). Haidt does not even endeavor to explain moral decision making in general – he tries to provide a theory of moral judgment (in the Other perspective) only. Haidt's approach in the 2001 paper is explicitly emotivist, in the sense that Haidt seems to defend the extreme position by David Hume according to which "we speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume 1960/1739, 415)[58]. The primary role of moral reasoning in the SIM is carrying out post-hoc rationalization. This is to say that moral reasoning concocts

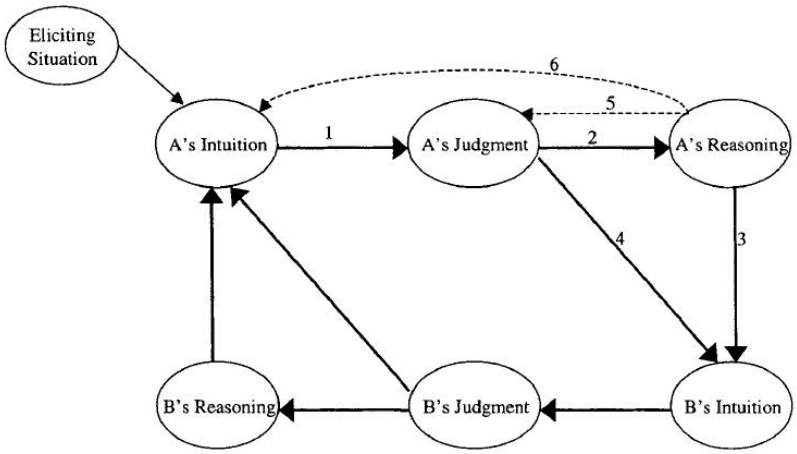
[58] Haidt (2012, 49) explicitly claims that Hume was right.

explanations of emotionally-driven processes in order to make them consistent with the subject's self-image. To use a comparison by Haidt, a moral reasoner is more like a lawyer that makes a case than like a judge that looks for justice: she tries to justify a judgment that has already been made at the emotive level[59]. Unbiased moral reasoning is rare: it only takes place when the subject has adequate time and processing capacity, a motivation to be accurate, no *a priori* judgment to defend, and when no relatedness or coherence motivations are triggered. This function of moral reasoning is similar to the activity of the "left-hemisphere interpreter" described by Gazzaniga (1998), a cognitive network located in the left hemisphere of the human brain which on the one hand fills the gaps in our perception of reality and renders it smooth and coherent, on the other hand confabulates just-so-stories for behaviors whose real causes cannot be acknowledged because they are destabilizing to the individual's self-image. According to a common reading of Haidt's paper, in the SIM moral reasoning has this function *only*. Most of the responses to Haidt (2001) were triggered by this interpretation of his claims[60]. Although it is true that some parts of Haidt's long paper seem to uphold extreme emotivist claims, this reading is in my opinion uncharitable and does not do justice to the SIM. Even though moral reasoning actually carries out the confabulating function described above, this is not the whole story about moral reasoning in the SIM. Haidt actually embraces a dual-process view, in which moral reasoning can be effective on moral judgments, regularly in social context and rarely

[59] Haidt (2012, 45) explains that he is not interested in a dichotomy between emotion and cognition, since emotions are cognitive processes in his view, but in the dichotomy between intuitions and reasoning, that are two distinct forms of cognition. However, Haidt has reached this level of clarity on this specific issue only recently.

[60] "Please don't forget the social part of the model, or you will think that we think that morality is just blind instinct, no smarter than lust. You will accuse us of denying any causal role for moral reasoning or for culture, and you will feel that our theory is a threat to human dignity, to the possibility of moral change, and to the notion that philosophers have any useful role to play in our moral lives […]" (Haidt and Bjorklund 2008a, 181)

at the individual level. The system is based on six links, that are depicted here (the figure is from Haidt 2001).



The most powerful links from the explanatory point of view are number 1, 2, 3, and 4 (full lines), whereas links 5 and 6 are weaker (dashed lines). Unreflective responses to the environment cause moral judgments to appear suddenly in consciousness – this appearance is a moral intuition, and this phenomenon is represented by link 1. As Dwyer (2009, 277) correctly points out, the link between moral intuition and moral judgment is not spelled out in a very clear way by Haidt. In particular, in the 2001 paper it was not particularly clear whether a moral intuition is a moral judgment or some emotional phenomenon that causes a moral judgment to appear. However, Haidt later modified the definition by making explicit that moral intuition is the appearance of an evaluative feeling: a moral intuition is the

> sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like–dislike, good–bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion (Haidt and Bjorklund 2008a, 188).

Moral intuitions are thus different from moral judgments. Be as it may, what matters here is that in the SIM emotional reactions to cases causally drive fast moral judgments. The emotions Haidt refers to are a subset of emotional phenomena known as the 'moral emotions'. Haidt defines them as "those emotions that are linked to the interests or welfare

either of society as a whole or at least of persons other than the judge or agent" (Haidt 2003a, 853). There are many moral emotions, but two sets are particularly important. The former is the so-called CAD set, Contempt, Anger, and Disgust, which are also known as the "other-critical" or "other-condemning" emotions, since they motivate punishment of moral violators or withdrawal from cooperation with them. The latter is the SEG set, Shame, Embarrassment, and Guilt, i.e. the so-called "self-critical" emotions, which promote cooperation and abidance by norms. The SEG emotions mainly aim at avoiding the CAD emotions of other members of the human group one lives in, so that the two sets are closely linked in a social context. These emotions are in turn triggered by specific events or situations. These sets of triggering events are innate, in the sense that they are "organized in advance of experience" (Haidt and Joseph 2007, 374) in the human brain. What these sets are is specified by the Moral Foundations Theory (MFT).

Link 2 represents the post hoc activity of reasoning. However, moral reasoning sometimes fails to provide convincing explanations for emotionally-driven moral judgments. In these cases, the subject cannot provide a justificatory reason for her moral judgment and while confronting the question "But why is behavior B morally wrong / praiseworthy / mandatory etc?" she finds herself unable to answer. This phenomenon is known as 'moral dumbfounding'. The typical example Haidt quotes is disgust towards incest. Humans do not like incest, even in cases in which there is no possibility of reproduction and when no physical or psychological harm can possibly be involved. Our dislike for sexual intercourse between people that share a sizable part of their genetic endowment (or that at any rate have been raised together) may be explained by biological evolution and could only justify aversion to cases of incest in which reproduction is possible. If reproduction is made impossible by contraceptive methods[61], there is no possibility to justify moral aversion to incest through biological evolution. Hence, experimental participants are usually unable to justify their moral responses in these cases, as instances of harm are equally excluded by

[61] In the hypothetical scenario Haidt (2001) depicts, not one but *two* contraceptive methods are used.

the scenario and both parties involved in the intercourse are adult and consenting. Importantly, this does *not* lead participants to revise their judgments, since the emotional responses that drove the judgment in the first place are still present and active.

Link 3 constitutes the biggest output of moral reasoning: modification of others' moral emotions, in particular through the reframing of scenarios. In the SIM, A's reasoning does not influence B's reasoning, but B's emotional responses to moral cases. The existence of Link 3, together with its being considered important by Haidt, shows that a reading of this model that focuses solely on post-hoc rationalization is unwarranted. An individual's moral emotions can be changed through another's reasoning, but also in another way. If a member of B's group or a friend of hers makes moral judgment J1, B is then induced to adopt J1. Similarly, an enemy of B's making moral judgment J2 pushes B to reject J2. This is represented by the social persuasion link, which bears number 4. Link 5 and 6 refer to private reasoning, that can have causal efficacy but, as we have said, under particular conditions only.

There are at least two interesting criticisms to the SIM: the former has been made by Pizarro and Bloom (2003), the latter by Fine (2006). Both try to show that the role of reasoning is not limited to post-hoc rationalization, but that reasoning plays an active role for moral judgment. As we have seen, Haidt largely concedes this *at the social level* by positing the existence of Link 3. However, these critics argue that moral reasoning is active *at the individual level* too – i.e. at the level of links 5 and 6 that Haidt deems to be rarely active. Pizarro and Bloom first notice a very important fact: intuitions can be shaped by processes that are rational and strictly cognitive even at the individual level. The most notable case is cognitive appraisal (or re-appraisal): the way in which a situation is construed (or re-construed) determines the kind of emotional reaction we have to it. To use the scenario proposed by Pizarro and Bloom, if you find a phone number scribbled on a piece of paper in a pocket in your spouse's jacket, you may get very angry if you already have suspects that he/she might have an affair. In contrast, in absence of these cognitive

premises, i.e. of this kind of appraisal, you might dismiss the finding as morally irrelevant – it might well be just a random phone number linked to a work issue. Furthermore, role taking (putting oneself in somebody else's shoes), which was one of the key component of the rationalistic model of moral development proposed by Lawrence Kohlberg (1969) and criticized by Haidt, is precisely, according to Pizarro and Bloom, a form of cognitive appraisal. Role taking elicits moral intuitions that are different from those that a moral judge would have in its absence, and often creates a positive feedback loop with empathy that can profoundly transform one's moral views. For example, you can contemplate the plight of a member of a minority you despise and that is unjustly discriminated against. By means of role taking and empathy with her suffering, you can profoundly change your moral opinions on the matter. But this process, Pizarro and Bloom argue, starts with reasoning and then involves a non-social emotion-reason interaction. Haidt (2003b) responds by biting the bullet and saying that moral intuitions are of course responsive to what we learn about our environment[62], yet this is relatively rare and happens mostly in social contexts. Haidt is making an empirical claim here, a claim that could in principle be tested and ascertained through an appropriate experimental setting. But it is not clear how to devise an experiment that shows how often role taking is used. It would be necessary both to keep the ecological conditions in the lab as close as possible to those present in standard social environments and to find a way to assess the use of moral reasoning vs. moral intuition in an implicit way, without availing oneself of self-reports, since those are

[62] It is useful to notice at this point that Haidt makes use of two different dichotomies. The former is between emotion and cognition in general, whereas the second is between intuition and reasoning. Intuition is meant to be a form of cognition as much as reasoning. Cf. for instance Haidt and Bjorklund (2008a, 200-201). In responding to Pizarro and Bloom, Haidt is claiming that intuitions have some cognitive component, i.e. are a part of cognition broadly construed. Therefore, they can react to changes in the environment and be influenced by role-taking under specific conditions. This might seem to be incompatible with a reading of moral intuition as a thoroughly emotive, non-cognitive phenomenon. Haidt seems to interpret moral intuitions in the latter way in some parts of his works, when he is more Humean than usual, but a more charitable reading of his position allows me to attribute him the view that moral intuitions have both a cognitive and an affective component. Of course this cognitive component is not a reasoning component. Furthermore, Haidt (2012, 45) explicitly maintains that intuitions are cognitive phenomena. The position by Prinz (2008), according to which moral judgments include the perception of an action and a sentiment, i.e. both a cognitive and an emotive component, seems to me to be more persuasive, provided that one has emotivist allegiances, than a strongly Humean reading of the SIM.

notoriously unreliable (cf. Nisbett and Wilson 1977). The empirical solution of this quandary seems thus not to be at hand.

The criticism by Fine draws on the one by Pizarro and Bloom, but actually manages to marshal empirical evidence to the effect that in some cases moral intuitions directly result from cognitive control processes. Cognitive control[63] is the capacity of reason to inhibit and override prepotent emotional or automatic responses. In a study, Monteith and colleagues (2002) asked participants to perform an Implicit Association Task (IAT)[64], a common procedure in experimental psychology that is *inter alia* carried out for the assessment of unconscious racial biases. Then participants were explained the structure and purpose of the task they had just carried out and were given their scores. Then subjects were asked to perform a second task, in which they were shown a list of words and had to tell on spot whether they liked them, using a standard Likert-like scale. Subjects that were feeling guilty because they showed a racial bias against Blacks in the previous IAT unconsciously inhibited their responses to first names that are typical to Black people and were significantly more likely than the other participants to say that they liked Black names. In this case, the inhibitory control of the racist bias transfers to new moral intuitions relative to new stimuli. Fine speculates that this kind of cognitive control can be rendered automatic, so that in the end a new, stable moral intuition is formed and the original racial bias is permanently suppressed in that individual[65]. Most behavioral patterns can be made automatic through repetition – learning to dance or to practice a martial art involves for example a similar kind of learning in the motor domain. I in turn speculate

[63] Also known as "executive function" and closely related to "working memory", i.e. short term memory used to store information that is necessary to perform a given task.

[64] The participant is presented with a series of words and names, and asked to categorize each word as 'pleasant'/'unpleasant', and each name as 'White name'/'Black name'. In the 'incongruent' condition, participants use the same keyboard key to categorize words as 'pleasant' and names as 'Black'. In the 'congruent' condition, the same key is used to categorize words as 'unpleasant' and names as 'Black.' If White participants are significantly slower at categorizing in the 'incongruent' condition, they demonstrate a negative affective association with Black names.

[65] After the publication of Fine's critique, Haidt explicitly admitted that this is possible: "[I]t is possible that some intuitions are just moral principles that were once learned consciously and now have become automatic." (Haidt and Bjorklund 2008a, 212)

that this could be one of the dynamics through which moral change takes place over time in human groups.

Recently, empirical evidence has been provided to the effect that cognitive reappraisal can dampen emotionally-driven moral intuitions[66]. Cognitive reappraisal includes a reinterpretation of emotion-inducing stimuli that reduces the impact of emotional experience. Hence, Feinberg and co-workers (2012) tested whether cognitive reappraisal has a significant influence on moral intuitions using the classic disgust-inducing scenario from Haidt, Koller, and Dias (1993): eating one's own dead dog. Their results corroborated the hypothesis. A greater spontaneous tendency to re-appraise was linked to fewer intuition-based judgments and participants made less emotion-based moral judgments when explicitly instructed by the researchers to re-appraise the emotions stemming from a movie clip they watched prior to the moral scenario.

Another recent claim on the same lines is the one by Campbell and Kumar (2012), according to which moral consistency reasoning, i.e. the requirement that similar cases must elicit the same kind of moral responses, can change moral intuitions in the long run. Let us suppose that cases C1 and C2 are similar, they elicit moral responses M1 and M2, the judge is more sure of her response to M1 than of her response to M2, and M1 and M2 are very different. This situation creates an inconsistency, which in turn arouses a negative moral response. This response conflicts with M2, which is weaker than M1, and contributes to change it over time. According to Campbell and Kumar, System1 is impenetrable in the sense that controlled cognition has no immediate effect on its internal operation and the outputs it yields. However, it does not follow that cognition has no long-term influence on System1. In short, while System1 is by definition *synchronically*

[66] Cognitive reappraisal can influence basic emotions too. For instance, suppose that Geoff is afraid of a dog because Geoff believes that (a) it will attack him, and in turn Geoff believes (a) because Geoff believes that (b) this dog is an American Pit Bull Terrier, (c) American Pit Bull Terriers regularly attack humans, and (d) American Pit Bull Terriers' attacks are dangerous to humans. If empirical evidence can be marshaled to debunk one among (b), (c), and (d), e.g. if Geoff understands that (e) this dog is a West Highland White Terrier, that his fear ought to disappear (where the 'ought' has to do with rationality). I thank Bernard Baertschi for pointing this out to me.

impenetrable, it may nevertheless be *diachronically* penetrable. However, Campbell's and Kumar's perspective is thoroughly philosophical and they do not feel compelled to strengthen their case with empirical evidence. What they write is plausible, but needs to be supported by experimental data.

Going back to Fine, she also has a precious theoretical insight. Consider a set of people P1 who hold moral judgment J3 and can provide reasons to justify it, as the judgment was in their case the outcome of a rational process. Later on, P1 can teach J3 to another set of people P2, who nonetheless cannot justify the judgment because they lack either the required knowledge or cognitive ability. However, this does not make J3 unjustified when uttered by people in P2. So people in P2, who are morally dumbfounded, are not necessarily holding unjustified judgments, nor are they maintaining claims that have an emotive origin. On the contrary, they may be making judgments that had a rational origin in P1 and were then learned by P2. Levy (2006b) expands this point adding an item: Haidt's emphasis on the social dimension of moral judgment. He correctly sees moral argument in human groups as a collective enterprise that is led by experts (who are not necessarily philosophers) and in which the work of moral justification is devolved to experts only. Moral believers would thus be rationally warranted to hold some moral beliefs even if they are victim of biases at the individual level: collective moral reflection among the experts may help to correct errors due to heuristics and biases. Hence, if Levy is correct, it is normal to find laymen who are morally dumbfounded: this is the regular outcome of the division of cognitive labor in society. Haidt's model should try to rule out this alternative explanation of moral dumbfounding. To my knowledge Haidt never answered to Fine's and Levy's remarks, but I might conjecture that he would answer that the phenomenon Fine and Levy underscore exists and that it rarely takes place.

The last point about the SIM concerns its empirical backing: what are the experiment that justify the SIM relative to competing views? First of all, there are Haidt's own experiments, for instance his cross-cultural examination of how people react to victimless

crimes, i.e. actions that harm none but that are considered wrong all the same[67], such as using your national flag as a rag to clean the toilet or eating your family dog once he is dead (Haidt, Koller, and Dias 1993). People who make moral condemnations against these actions are morally dumbfounded and their judgments are better predicted by their self-reports concerning their emotional involvement than by their self-reports about the alleged presence of harm. Then, Haidt (2001) quotes a significant amount of empirical work but, as Levy (2006b) correctly points out, it is mostly domain-general, i.e. not linked to the moral domain in any specific way. Hence, it is not decisive to establish the validity of a theory in moral psychology, because the generalizability of results from experimental psychology broadly construed to the specific and normative domain of moral judgments must specifically be tested.

The stronger empirical results in favor of the SIM have appeared after the publication of the original article. For example, Wheatley and Haidt (2005) showed that the hypnotic triggering of a pang of disgust in experimental participants made moral judgments more severe. Experiment 2 in this article also features the so-called Student Council case, a case in which moral violations are simply absent[68]. In absence of hypnotic disgust, participants rated that this case was not morally problematic. In presence of hypnotic disgust, participants said that the case was less uncontroversial than before, but still approved of the proposed action. This experiment hence does not show, as it has been suggested (for example by Prinz 2006), that *emotions are sufficient for moral judgments,* but it definitely shows that emotions *causally modulate moral judgments* in a statistically significant way. Of course, this has nothing to do with issues concerning the justification of such a

[67] Many ethical systems, such as Roman Catholic ethics and Kantian ethics, include the belief that there are crimes that require no victim at all (e.g. consensual homosexual anal intercourse for Roman Catholic ethics and lying or masturbation for Kantian ethics). Hence, "victimless crime" sounds as a paradoxical expression only if one starts from a Turiel-like point of view concerning the proper domain of morality. This has no repercussion on Haidt et al's experiments, though. These experiments were just mapping who "moralizes" 'victimless crimes' and who does not.

[68] This is the precise wording: "Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He [tries to take/often picks] topics that appeal to both professors and students in order to stimulate discussion.'' (Wheatley and Haidt 2005, 782). It is admittedly difficult to see how this could involve moral problems.

judgment, i.e. emotion *per se* is unable to provide any kind of adequate justification[69]. Other results show modulation of moral judgments by emotions. For instance, Valdesolo and DeSteno (2006), whom I already mentioned in Ch. 2, showed that induction of positive emotions make moral judgments more lenient. Then, in a series of famous experiments, including the "fart spray" and "filthy desk" set ups, Schnall and coworkers (Schnall et al. 2008; Schnall, Benton, and Harvey 2008) have shown that manipulations of the experimental environment, such as letting the participant sit at a filthy disk or perceive a strong stink, can make moral judgments more or less harsh. In particular, moral judgments were harsher when made near a purse that had been sprayed with a stinky gas or when made by a participant with a high private body consciousness[70] and who at the same time sat at a filthy desk[71]. On the contrary, priming for the concept of cleanliness using a scrambled-sentences task[72] made moral judgments less harsh. To add to this body of evidence, sentimentalist philosopher Jesse Prinz and his co-workers (Eskine, Kacinik, and Prinz 2011) have shown that bad tastes influence moral judgments in the same way as nasty smells.

These results show that emotions pervasively modulate moral judgments, but they *do not show*, so far, that emotions are *sufficient* for moral judgments.

The SIM is quite a flexible model and, as Haidt says, the emotional dog can easily learn new tricks. Therefore, there is no strong counter-evidence against the SIM at the moment, as Haidt can almost always answer "Yes, this phenomenon you have spotted exists, but it's rare." The most serious piece of counter-evidence is the experiment conducted by Paxton, Ungar, and Greene using the CRT. It shows that the manipulation of reflectivity at the

[69] Kass (1997) would disagree on disgust not providing justification, but I will not discuss Kass's claim here, as it would bring us too far away. For positions opposed to those of Kass, cf. Nussbaum (2004).

[70] Private body consciousness is "people's general attention to internal physical states" (Schnall et al. 2008, 1100).

[71] This is the description of the filthy desk: "An old chair with a torn and dirty cushion was placed in front of a desk that had various stains and was sticky. On the desk there was a transparent plastic cup with the dried up remnants of a smoothie and a pen that was chewed up. Next to the desk was a trash can overflowing with garbage including greasy pizza boxes and dirty-looking tissues" (Schnall et al. 2008, 1101).

[72] Participants are given a set of 5 or 6 words and have to compose a meaningful sentence with them. The unconscious priming is created by putting a lot of words from a particular semantic area (e.g. "soap", "wash", "water", "vacuum cleaner", "pure", "hygiene" etc for cleanliness) into the word pool for the task.

individual level influences moral judgments. Haidt's standard reply seems to be available here, even though it is not particularly convincing. Furthermore, the SIM does not explain all the data either and it is admittedly silent on results at the neural level. The SIM is plausible, is supported by some empirical data, yet it is unable to deliver a knock-out strike to alternative hypothesis and has been subject to intensive, intelligent criticism.

As to the MFT, it describes extent and structure of the moral domain and not directly the way in which moral judgments are formed. Therefore, it is rather orthogonal to Greene's model. However, the MFT importantly complements the SIM, although it is independent from it. Haidt claims that there are five (or six, according to the latest news) roots of morality that correspond to related cognitive structures. According to the MFT, the pangs of emotion that bring moral judgments about are due to the activity of these foundations. So the MFT describes what generates moral intuitions in the Haidtian sense: this is the point of interaction with the SIM. The MFT is a nativist theory, i.e. a theory of the mental and neural bases of moral judgments that posits more innate components than a general, multi-purpose learning capacity. "Innate" is a very unclear and controversial term and I will not delve into details here[73]. However, Haidt means here that humans are "prepared" to develop emotional susceptibility to a set of stimuli, which are mostly actions and omissions. The development of this capacity for emotional reaction is rather robust in ontogeny, even though environmental input is required for it to develop and counts as a necessary condition for the capacity to arise. In this way social learning can modulate the details of this capacity for emotional reaction. Haidt introduced the MFT in Haidt and Joseph (2004) and then developed it in Haidt (2007), Haidt and Graham (2007), Haidt and Joseph (2007), and Graham et al. (2011). The latest version of the MFT is in Haidt (2012, Part 2). The MFT has important forebears. First, the aforementioned ethnological study in Orissa – India by Shweder et al. (1997) highlighted a structure of the moral domain composed of three clusters: Ethics of Autonomy, Ethics of Community, Ethics of Divinity.

[73] In order to have more details, cf. the enlightening paper by Mameli (2008).

The Ethics of Autonomy includes moral concerns about physical harm, psychological harm, and distributive justice. It largely corresponds with Western liberal morality and with the aforementioned description of moral judgments given by Turiel (1983, 3) as "prescriptive judgments of justice, rights, and welfare pertaining to how people ought to relate to each other." Ethics of Community has to do with respect for hierarchy, loyalty, obligations towards one's in-group, and so on. The violator is mostly seen as untrustworthy, treacherous, and cannot be considered a full member of the group. Ethics of Divinity has to with religious purity, and sex and food taboos. An upholder of an ethic of divinity sees the body as a temple that must be kept pure and would see violators of purity norms as breeching the boundary that separates humans from non-humans, i.e. losing their full human status to acquire some of the baseness of the beast. Rozin et al. (1999) suggest an interesting mapping between these three ethical codes and the moral emotions that belong to the other-critical triad: Contempt, Anger, and Disgust. Violations of the Ethics of Autonomy elicit Anger, violations of the Ethics of Community evoke Contempt and shunning, whereas violations of the Ethics of Divinity mostly elicit Disgust. Since the initials of the ethics and the emotions happen to correspond, this was dubbed as the CAD Hypothesis. In Haidt and Joseph (2004) four foundations are proposed, with the Ethics of Autonomy getting split into Suffering and Reciprocity. The five roots scheme, that has lasted for at least five years, includes (1) Harm/Care, (2) Fairness/Reciprocity, (3) Ingroup/Loyalty, (4) Authority/Respect, (5) Purity/Sanctity. These roots evolved in order to solve five adaptive challenges in the EEA: (1) caring for vulnerable children, (2) forming partnerships with non-kin to reap the benefits of reciprocity, (3) forming coalitions to compete with other coalitions, (4) negotiating status hierarchies, and (5) keeping oneself and one's kin free from parasites and pathogens (Haidt 2012, 125). Haidt is currently adding a six root, Liberty, that represents the value humans usually attach to individual freedom. It was added to explain the morality of US Libertarians. Many criticisms could be leveled to this description of the moral domain. The set of roots or foundations is plausible

but seems to be arbitrary. In particular, it is not clear how fine-grained the Foundations ought to be. The results published by Graham et al. (2011, 375-376) show that there is an important gap between the first two foundations that came from the Ethic of Autonomy and all the others, thereby mirroring the theory Haidt has put forward to explain the distinct moral beliefs of conservative and liberals in the Western world (Haidt and Graham 2007; Haidt, Graham, and Joseph 2009; Graham, Haidt, and Nosek 2009). Haidt claims that liberals avail themselves of only two of the five existing Foundations, whereas conservatives have a richer morality that takes all of the five Foundations into account, thereby experiencing frequent trade-offs. So, it is unclear why there should be five Foundations instead of two, one that stands for Liberal ethics and another that indicates non-Liberal ethics. Alternatively, we might move from five to more foundations, since we might discover some moral judgments that fit only with difficulty into the current framework, as the case of liberty demonstrates. Suhler and Churchland (2011) quote other two candidates to inclusion: industry and modesty. Although modesty could be easily put into the Authority/Respect root, industry seems to be more independent, although it shows some ties with virtues in the Ingroup/Loyalty group. However, Suhler and Churchland successfully show that the taxonomy proposed by Haidt and co-workers is rather contrived. On the other hand, Haidt himself does not insist much on the exact number of the foundations, making this objection not very pointed. Indeed, he explicitly admits that the exact number of the Foundations might be different from five[74]. A second critique by Suhler and Churchland may have more bite. It concerns modularity. The standard definition of a module comes from Fodor (1983) and according to it a module is an innate, fast, informationally encapsulated, functionally specialized computational mechanism. Informational encapsulation refers to the idea that there is no or little transmission of information between the module and other neural/mental structures, i.e. the module has no

[74] "We do not claim that there are *only* five foundations. There are probably many more, but we believe the five we have identified are the most important ones for explaining human morality and moral diversity." (Haidt and Joseph 2007, 385).

or just a few connections with other parts of the cognitive system, except at the input and at the output levels. Its internal workings are thus not influenced by other parts of the mind/brain. According to this stringent definition, there are just a few modules in the mind/brain. Nonetheless, drawing on the work of anthropologist Dan Sperber (1996), Haidt uses a much broader definition of "module". According to Sperber (and therefore Haidt), modules are highly variable (some meet all of Fodor's criteria, some meet only a few), they are often nested within each other, and are usually not innate: they are generated during development by a smaller set of "learning modules" which, in turn, are innate. So Haidt and Joseph (2007) maintain that the moral foundations must be seen as "learning modules" that produce a mass of second-order modules ('teeming modularity'). Second-order modules are in charge of the formation of moral intuitions by coupling the perception of some features of a situation (usually being an instance of a virtue or a vice) with a moral emotion, such as a CAD emotion or a SEG emotion. Suhler and Churchland (2011, 2105) object that this view of modularity is very expansive and risks becoming vacuous. However, Haidt's proposal is relatively moderate in comparison with the idea of massive modularity brought forward for instance by Pinker (1997). There is no way Haidt's position can be considered untenable if one considers the very diverse positions that have appeared in the modularity debate inside cognitive science. But Suhler's and Churchland's attack is more precise than this: they charge Haidt with providing no operational criteria to discriminate between the cognitive processes that are innately prepared in the brain and those cognitive processes that are not so prepared[75]. Furthermore, Suhler and Churchland claim that second-order modules seem to be "little more than a way of designating somewhat arbitrarily chosen (and, perhaps, arbitrarily fine-grained) stimulus–behavior patterns without shedding any light on the underlying processes' computational workings" (2011, 2106). Finally, they claim that the recursive architecture of the brain, in which spontaneous activity is rampant and circuits are loopy more often than not makes the idea

[75] "But how do we know when a trait emerges from a learning module and when it does not?" (Suhler and Churchland 2011, 2105)

of neural modules rather far-fetched. Although the neuroscientific arguments marshaled by Suhler and Churchland to attack the idea of teeming modularity seem to be rather orthogonal with psychological modules such as Haidt's, which could have multifarious neural instantiations, I agree with them on this: Haidt's explanation of how his modules work remains quite generic. However, Haidt and Joseph (2011) claim that they are not required to give specific details on the computations that are carried out by the modules, and even less about the neural correlates of the modules (both learning and second-order). They even make the bolder claim that, at the state of the art of neuroscience, it is impossible that some psychological phenomenon, such as the informational encapsulation of moral dumbfounding[76], can be debunked by neuroscientific results or dismissed because non-consilient with them. The latter claim is rather controversial. As Haidt and Joseph (2011) correctly notice, there are different traditions in cognitive science. Scientists that come from East Coast universities in the US tend to be nativist and to widely apply the notion of 'module', whereas West-coast empiricists are skeptic toward modularity and give more emphasis to general-purpose learning mechanisms. Their favored strategy of explanation is social learning. Suhler and Churchland come from the West, Haidt and co-workers come from the East. So, West Coast cognitive scientists are unlikely to buy Haidt's claim for the psychological irrelevance of present neuroscience, since they have very high standards for attributing the status of 'module' to a mental function, and this standards include neuroscientific specifications. In contrast, Haidt and Joseph stress that their modules ought to be seen as functional (as opposed to neuro-anatomical) ones and that the requests made by Suhler and Churchland to provide neuroscientific and computational details are unusual in experimental moral psychology[77]. Nonetheless, it is hard to tell whether this request is unreasonable or uncommon, as there is no central

[76] Moral dumbfounding is considered a form of informational encapsulation because unjustified moral judgments are not modified even in front of the evident impossibility of justification. The result of other cognitive processes (e.g. those that are in charge with moral justification) does not then influence the output of moral modules, i.e. moral emotions and corresponding intuitions.

[77] "We are surprised to hear that this is now a common expectation." (Haidt and Joseph 2011, 2218)

authority that establishes what kind of features a theory in experimental moral psychology must have in order to count as a "good theory." As for the notorious "burden of proof" in philosophical arguments, it seems that this kind of charge can be made quite arbitrarily by anybody against anybody else, since methodologies in the cognitive sciences are far less clear-cut than in physics or in molecular biology[78]. Of course it would be better for the MFT to feature Suhler's and Churchland's *desiderata* (and this is not controversial), but it is not obvious whether Haidt and co-workers *must* necessarily deliver these goods now in order to insure the survival of MFT. Perhaps the MFT might be seen as "good enough" even without details about computational mechanisms and neuro-anatomical correlates. At which height the bar must be placed seems to be largely arbitrary and to depend on the above-mentioned cultural affiliations in cognitive science. Hence, even the point about modularity and computational details fails to deliver a knock-out blow to the MFT. At the moment the MFT is a plausible hypothesis that is supported by some empirical results. However, it needs to address some problems of vagueness and to gain in computational detail in order to increase its credibility.

3.2. Moll's EFEC model

Brazilian neuroscientist Jorge Moll has been one of the fathers of neuroscience of morality, together with Joshua Greene (e.g. Moll et al. 2002a, 2002b). Nevertheless, his model of moral cognition is very different from Greene's. Contrary to Greene's dual-process model, it is a single-process model. A single-process model hypothesizes that all moral judgments result from a single stream of mental operations, without conflicts between systems. In the view expounded by Moll and his co-workers (2005, 2008a, 2008b), moral judgments and moral emotions[79] are caused by EFECs (Event-Feature-Emotion Complexes).

[78] Happily enough for cognitive scientists and experimental psychologists, methodology in cognitive science is clearer than philosophical methodology.
[79] Moll defines in the same way as Haidt (2003).

An example may be useful to give the reader an idea of how this model works. I quote it from Moll et al. (2005). In front of an orphan girl who stares sadly at us and who has little chances of being adopted because too old, we form three representations: an Event, a Feature, and an Emotion. The Pre-Frontal Cortex (henceforth PFC) provides the Event representation: "The girl is an orphan and the odds of adoption are low." The posterior part of superior temporal sulcus (henceforth p-STS) and the Anterior Temporal Lobe (henceforth ATL) contribute the child's sad facial expression and the concepts of "helplessness" and "sadness", i.e. the Features. Lastly, the so-called limbic system, which is mostly sub-cortical, yields central motive states (feeling emphatic sadness, anxiety, and attachment), i.e. the Emotion. These three components generate a moral emotion of compassion toward the girl and a moral judgment that it would be morally praiseworthy to help her.

As to the details of the model, the Event is a representation of some action or state of affairs. One of the characteristic of Moll's model is its rootedness in neuroscientific data. Hence, each of the three components of an EFEC is instantiated by a specific brain area, as I have shown in the example. The Event cognitive component is computed by the PFC, i.e. the pre-frontal cortex, the most specifically human area of the brain, since it is the one that evolved last. Moll relies on the interpretation of the PFC's activity provided by his co-worker Jordan Grafman (1995), i.e. on the so-called Structured-Event-Complex (SEC) framework. The SEC framework maintains that the PFC carries out its cognitive control function over behavior through "long-term memories of event sequences that guide the perception and execution of goal-oriented activities, such as going to a concert or giving a dinner party" (Moll et al. 2005, 803). In other words, the PFC forces behavior to follow one of these representations, as though it were a rule. Furthermore, different parts of the PFC are allegedly in charge with distinct representations of this sort. Namely, the VMPFC is linked to social and emotional complexes, the anterior and lateral PFC with new representations or with representations that include branching outcomes and probability

assessment, and more posterior parts of the PFC with SECs that have already been learned. Hence, Moll hypothesizes that lesions to different parts of the PFC should cause differential deficits and in this he is backed, at least partially, by empirical evidence, such as the studies by Damasio (1994) on VMPFC lesions, both with late onset, as in Gage's case, and with early onset (cf. Anderson et al. 1999).

As to the second component of an EFEC, the Feature, it refers to context-independent social concepts, such as "sadness", "helplessness", "tactlessness", "honorability", etc. These social concepts must be read by the individual through the other's verbal, postural, or facial expression. The p-STS and the nearby and partially overlapping Temporal-Parietal Junction (TPJ) are, together with other regions[80], in charge with mindreading and extracting this information from visual and auditory stimuli. In contrast, the concepts themselves are stored in the ATL, and especially in the temporal pole. It is well known that patients with Fronto-Temporal Dementia (FTD), a disease that deeply affects ATL neurons, exhibit important deficits in social behavior (cf. Mendez et al. 2005; Neary et al. 1998) and that temporal poles are involved in the neural processing of representations of harm to other human beings (Heekeren et al. 2005). Furthermore, Zahn and coworkers (2007) have experimentally shown that abstract social concepts are correlated with activation in this brain area. As to the TPJ, there is plenty of evidence for its importance in general mind reading, i.e. attribution of mental states to others (Decety and Lamm 2007; Saxe and Kanwisher 2003; Saxe and Wexler 2005) and in the attribution of intentions and beliefs in moral judgments more specifically. To this effect, Liane Young and co-workers have on the one hand demonstrated through fMRI that r-TPJ BOLD signaling significantly correlates with participants reading descriptions of morally relevant actions in comparison with irrelevant actions (Young and Saxe 2009), on the other hand they have shown that rTMS on the r-TPJ specifically reduces the importance of the assessment of intentions on

[80] Notably the MPFC. Cf. Amodio and Frith (2006).

moral judgments, bringing about evaluations that mostly focus on outcomes (Young et al. 2010).

As to the third component, Emotion, it is described by Moll as providing to moral judgments their motivational force. If the moral machinery were disconnected from emotional components that are embedded in sub-cortical areas such as the striatum, the tegmental area, the bilateral amygdala, and others, moral judgments could not influence behavior. Moll differentiates between emotions and "central motive states", such as "undirected anxiety." Emotions are, in Moll's opinion, the result of a coupling between a stimulus and a central motive state, that must hence be seen as more basic than emotions themselves. So, what gets associated with an Event and a Feature is not an emotion *sensu stricto*, but a general affective state that gives to the complex its motivational drive.

The moral domain Moll takes into account is very broad. The precise definition he gives is "the sets of customs and values that are embraced by a cultural group to guide social conduct" (Moll et al. 2005, 799). This definition seems to include positive laws and conventional rules, that are not universally regarded as having moral import. This expanded moral domain does not encompass moral judgments about cases only, but also moral emotions in general, so that phenomena such as pride for victory, embarrassment for the violation of a convention, and compassion towards the suffering of the victim of a natural disaster all fall into the purview of morality according to Moll's view. Hynes (2008) indeed accuses Moll of conflating moral emotions and social emotions. In front of this attack, Moll and co-workers (2008c) bite the bullet and maintain that there are so far no empirical data that back the distinction between social and moral emotions. This gives a sense of how big the domain of morality is in Moll's view. Indeed, Moll et al. (2008a) claim that their view is compatible with Haidt's MFT, another example of extended moral domain. Nonetheless, Moll and his colleagues (2008a, 162) try at the same time to specify the difference between the social and the moral domain by claiming that moral cognition is specifically linked to altruistic actions, but some foundations of morality highlighted by the

MFT, such as purity, seem to have little to do with altruism. The Jewish prohibition to eat crustaceans, for instance, seems to me to be uncoupled from altruistic motivations or helping behavior. At this point, Moll cannot have it both ways: either he concedes that his link between altruism and morality cannot be a double conditional, or he concedes that food taboos are not moral issues, thereby jettisoning the MFT.

That being said, the main point that interests me here is how Moll's theory differs from Greene's. Whereas Greene posits a potential (albeit rare) conflict between System1 and System2, Moll's idea is that moral emotions always include affective and rational components, so that there can be no real conflict[81]. As to moral dilemmas, which are deemed by Greene fruitful epistemic tools able to flesh out the "fault lines" of the moral domain, Moll, following Haidt and Kesebir (2010, 807-808), Casebeer (2003), and others, thinks that they lack ecological validity and are therefore bad epistemic tools whose use contributes to create a distorted view of how the moral domain actually works. If there are conflicts in the moral domain, these are extremely rare and are due to the fact that the moral circuit produces two different moral emotions relative to a given situation. In any case, there is no possibility of conflict between reason and emotion *per se*, but between two complexes that are to the same extent both emotional and rational. In the light of all this, Moll claims that the concepts on which Greene bases his analysis — 'personal', 'impersonal', 'deontology', 'utilitarianism' — ought to be analyzed into "clear cognitive components" (Moll et al. 2005, 801). Finally, there are also disagreements between Moll and Greene about the neural correlates of some mental functions. For instance, Greene takes the DLPFC (the 'accounting department') to be the neural basis of cost-benefit analysis, whereas Moll thinks that this function is carried out by both the anterior parts of the PFC and the limbic system, which alone can attribute value to choices. Hence, in Moll's view, this function is both rational and emotional, whereas in Greene's view it is a

[81] It should be reminded, however, that Greene et al. (2004) concede that all moral judgments have an affective component and that Greene (2008a) tries to insert emotions in consequentialism through the concept of "currency-like emotions".

clear example of rational function and perhaps even the rational function *par excellence*. Moll's model does not fit well with all data points, since it contains an interpretation of VMPFC function that has been proven false on experimental grounds[82]. However, it fits non-trolley experimental results quite well and makes us understand how experimental moral psychology is still a very open game. Since it is a very different model from Greene's, its viability shows that Greene's descriptive idea cannot at the moment be taken for granted as the empirical "truth" about human moral behavior. Both Haidt's theory and Moll's deem emotions key causal antecedents of moral judgments. On the contrary, the view I examine in the next section sees emotions as outcomes of moral judgments, not as causal factors.

3.3. Universal Moral Grammar and Gazzaniga's neuromoral theory

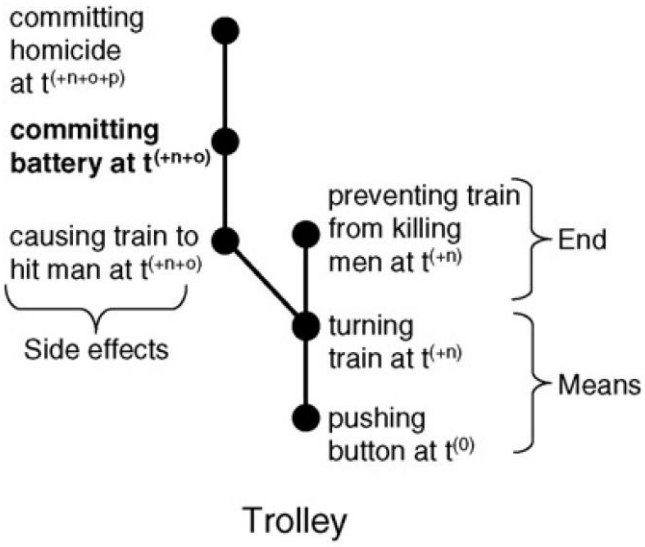
Both Haidt and Moll are to a certain extent emotivists, theorists that think that moral judgments are, *inter alia* or *in toto*, caused by emotive states. Hence, one of their targets are the old theories by Kohlberg according to which moral judgments are mostly caused by explicit, discursive reasoning. Apart from Greene's view, there are several ways out from the Kohlbergian paradigm: one is the strong emotivism championed by Haidt, another is the single-process model by Moll. Here I explore a fourth one, that is the idea of a moral grammar. The idea was originally brought forward by John Rawls himself (1999/1971, § 9). As the human sense for linguistic grammaticalness requires, in order to be explained along the lines of Chomsky (cf. for instance 1965), cognitive constructs that far outstrip in complexity the explicit content of the individual's grammatical knowledge (i.e. folk grammatical concepts), so the explanation of the human sense for moral violations requires cognitive constructs that are much more complex than our folk moral concepts. The main advocates of this line of research in experimental moral psychology are philosopher Susan Dwyer (cf. especially 2009), legal theorist John Mikhail (2007, 2008, 2011), and

[82] I mentioned this above in Ch. 2, when referring to the Moretti paper.

psychologist Marc Hauser (2006; Hauser, Young, and Cushman 2008). Hauser's career as a scientist ended abruptly in 2010 after a three-year internal investigation at Harvard[83] ascertained that he was responsible for eight counts of scientific misconduct, including data forgery. However, none of the counts involves experiments on humans in experimental moral psychology, so that I will consider valid the work done by his lab at Harvard on this topic. Furthermore, I deal here more with the theoretical positions by Hauser than with the experimental results from his lab. The advocates of the Universal Moral Grammar (UMG) differentiate themselves from the Haidtian mainstream by positing that moral judgments are the outcome of computations that are cognitive *sensu stricto*, i.e. devoid of affective content, and implicit, so that the subject is not aware of them. These computations are both unconscious and very fast. Hence, these processes are a direct challenge to Greene's dual-process-model because they include some traits that are typical of System1 and some traits that are more akin to System2, thus blurring the demarcation dual-process theorists want to draw in human cognition. The theorist that more consequentially spells out the structure of the alleged grammar humans use to make moral judgments is Mikhail (cf. 2007, 2008). First, I examine the structure of UMG according to Mikhail's version. Secondly, I discuss the role of emotions in this model, mostly taking into account Huebner, Dwyer, and Hauser (2009). Thirdly, I examine the proposal by these theorists that the cognitive bases of moral judgments form a domain-specific network, or even (on a more extreme formulation) a module in the Fodorian sense. UMG is a theory at the computational level, so that UMG is not interested in the neuro-physiological underpinnings of the computations that are identified by the theory. In contrast, Moll's theory is for instance active both at the computational and at the neural level, as well as Greene's dual-process view. Hence, it will not be possible to debunk such a theory by using neuro-scientific contents, unless one wants to assume with Suhler and Churchland (2011) that psychological theories have to provide neuro-physiological details on pain of

[83] This investigation was prompted by a report from one of his closer co-workers.

being declared invalid. The moral module ultimately operates as a function whose domain are instances of actions and omissions and codomain are deontic statuses. The moral module first takes a given action and parses it into a graph-like representation, like this one (picture from Mikhail 2007, 146). This graph represents the choice of turning the trolley to the side-track in the standard *Switch* dilemma. The module carries out this transformation (from action to graph) through specific conversion rules.



Then a given structural representation yields a deontic status (such as mandatory, forbidden, non-mandatory, and non-forbidden, i.e. merely allowed — see box 2 in Mikhail 2007, 144) through deontic rules. Hence, the work of the moral psychologist in this case is to discover both conversion rules and deontic rules. Both are often unconscious. Deontic rules correspond to moral principles, such as the Doctrine of Double Effect (henceforth DDE) and the DDA[84]. Cushman, Young, and Hauser (2006) have shown that these two principles, together with the Contact principle that resembles to Greene's 'personal force' factor, are active in moral judgments, but they have different level of availability to the conscious mind. The DDE acts in an extremely covert way, whereas most subjects who avail themselves of the DDA can explicitly quote it as a justifying reason of their judgments. As to deontic rules, Mikhail focuses on two: the prohibition of purposeful

[84] The DDE is the idea that one is more justified at inflicting physical harm if the harm is a foreseen side-effect of an action than if the harm is a necessary means to reach the end of the action. The DDE can also be applied to morally blameworthy actions that are different from inflicting physical harm, i.e. it is not specific for harm. Cf. Quinn (1989b).

battery and the DDE. In his opinion, these two intuitions can account for most of our responses to physical violence, trolley-like cases included. As to conversion rules, they first examine the temporal structure of the event, then add a causal structure, identify good and bad outcomes, and then superimpose an intentional structure of means, ends, and side-effects to the morally-evaluated causal pattern. In other words, the following features are progressively added to the raw stimulus, in this order: temporal features, causal features, moral features, intentional features. The final result is one of the graph-like representations above. Details need not interest us here: an overall view of Mikhail's proposal on this is sufficient. What is most interesting is that conversion rules, in Mikhail's opinion, are one instantiation of the poverty of the stimulus problem that prompted Chomsky to put forward transformational-generative grammar. Mikhail argues that causal patterns and intentional patterns are pieces of information that cannot be found in the stimulus and there is no way the subject can learn them from the environment. Hence, these notions (cause, effect, means, end, etc) need to be innate, in a sense of "preparedness" that is similar to the one that applied to Haidt's MFT. So this theory, like Haidt's, Moll's, and Greene's, is nativist, in the sense that it denies that morality can arise from a general learning mechanism coupled with environmental stimuli. Contrary to other nativist theories, however, UMG tends to favor simplification and indeed it does not attack trolley cases for being devoid of ecological validity. On the contrary, Mikhail is a fan of trolley-like cases since they allow him to show the essential structure of the moral module, to highlight the fault lines of the cognitive function of interest. Thus, as to this point, Mikhail and Greene are in perfect consilience. And this is enough as to the structure of the UMG.

As we have seen, emotions play little or no role in the computations described above. However, it cannot be denied that emotions have something to do with morality, especially after the experiments by Wheatley and Haidt, Schnall, and others have been published. So, the moral grammarians need to explain us what to do with emotions. This kind of explanation is provided by Huebner, Dwyer, and Hauser (2009). Their line of defense

involves diverse facets. First, they posit a distinction between these positions: "emotion accompanies moral judgments" vs. "emotion constitutes moral judgments," where they defend the former. This distinction looks unclear to me. From the experimental point of view, it cannot happen by chance that emotions accompany moral judgments in a statistically significant way. Therefore, there must be some sort of causal connection between the two. This connection could even be mediated by a third variable that causes both moral judgment and emotions, but some sort of causal story must be in place. Hence, to say that 'emotions accompany moral judgments' means one of these three:

(1) Emotions are one of the causes of moral judgments;

(2) Emotions are one of the effects of moral judgments;

(3) Emotions and moral judgments depend on some common upstream cause.

What does it mean that emotions 'constitute' moral judgments? If this means that emotions are necessary and sufficient conditions for a moral judgment to be performed, as for example Prinz (2006) claims, this is a qualified and stricter version of (1). The moral grammarians seems to uphold (2) instead. They argue that emotions are not part of the computation that maps actions and omissions onto deontic values, that emotions are an effect of these computations, and that they provide individuals with motivations to modify their behavior in the light of the attribution of deontic value that has been performed by the moral module. This position, which is dubbed as the 'pure Rawlsian' position in the Huebner, Dwyer, and Hauser paper, stands in contrast with experimental results, especially with the ones showing emotional modulation of moral judgments (Wheatley and Haidt 2005, Schnall et al 2008, etc. Cf. § 3.1.). So the authors seem in the end to favor instead a 'hybrid Rawlsian' position, according to which emotion causes moral judgments together with the output of the moral module, i.e. attribution of deontic values. Moreover, they claim that emotions could act on moral judgments via redeployment of attentional resources, i.e. prompting attention to some aspects of the scenario that are morally relevant for independent reasons. Nonetheless, there is no positive empirical evidence, at least to

my knowledge, to buttress this idea, however intuitively plausible it may be. On the other hand, there is no empirical evidence that emotions are completely sufficient for moral judgments, although it must be admitted that emotional modulation is powerful and widespread. The study by Wheatley and Haidt (2005) is unfortunately unable to show that disgust can create moral condemnation out of nothing, so that the idea that emotion is making some scenario traits more salient cannot be ruled out at the moment. So it must be concluded that the jury is still out. If somebody provided conclusive empirical evidence that emotions are sufficient for moral judgments, then the 'hybrid Rawlsian' model would be empirically discredited, but we are not there yet and perhaps we will never be. It is sure that emotions exert a *causal* role in the making of moral judgments, but we do not know at the moment whether they are *sufficient* causes.

Thirdly, the moral grammarians think that the moral module is domain-specific. On a 'hybrid Rawlsian' view, this circuitry is unique in the way "the attribution of intentions and goals connects with emotions to create moral judgments of right and wrong" (Hauser 2006). Thus, this circuit works for moral judgments only. Domain-specificity is hotly contested in the neuroscience of morality. Both Greene and Haidt (2002) and Young and Dungan (2012) argue that moral processing is nothing specific, but resides in applying to particular stimuli (i.e. morally relevant stimuli, however defined) cognitive resources that are *per se* multi-purpose. Greene and Haidt claim that "If one attempts to 'deconfound' moral judgment with everything that is not specific to moral judgment (emotion, theory of mind, mental imagery, abstract reasoning, and so on) there will almost certainly be nothing left" (2002, 523). However, the point is very difficult to settle. In order to assess a claim for specificity, we ought to know whether this module performs other functions. Unless the neural correlates of the module are explored, it is very difficult to assess such a position. Even going to the neural level (which is precisely what Mikhail and likely-minded scholars refuse to do) and discovering that, for the sake of argument, the module is instantiated by a network of four given regions in the brain, the idea of domain-specificity seems to be

rather difficult to corroborate or refute. What we currently know about the human brain is by far too little to know exactly what the mapping between psychological processes and functions on the one hand and networks of brain regions on the other hand is. As I have said in the preceding discussion on reverse inference (§ 2.2.), we still have troubles to define both a psychological ontology of mental functions and a neuro-anatomical ontology of brain regions. Hence, this dispute cannot be solved now, provided that the domain-specific moral circuit is seen as a complex network of brain regions and not as a single, continuous brain area. Indeed, if the module were a single patch of neural tissue, we would know that in the overwhelming majority of cases it is involved in more than one function. As I have argued above, the case for one-to-one mapping between mental functions and brain regions is exceedingly hard to make.

Apart from my precedent objections to the idea of a UMG, there are other three pieces of criticism that should be addressed: one by Joshua Greene (2008b), another by French cognitive scientists Dupoux and Jacob (2007), and a last one by philosopher of biology Kim Sterelny (2010).

Greene notices that Mikhail accepts "trolleyological" evidence and asks: "Please explain the moral dilemmas data using your UMG." In particular, Greene asks Mikhail to explain why, assuming (as Mikhail does) that the only active rules for moral choice are a condemnation of battery and the DDE, (a) some people decide to throw the bystander down the bridge, (b) in the *Loop* case[85], half of the sample behave in a way that is the opposite of what UMG predicts, i.e. UMG predicts that people shouldn't turn the trolley on the side-track, but half of the sample does that anyway. Furthermore, Greene asks to explain the results by Wheatley and Haidt, Valdesolo and DeSteno that show modulation

[85] The *Loop* case (also known as *Ned*) is a variant of *Switch* invented by Thomson (1985) as a counter-example to the DDE. In the *Loop* case, the death of the man on the side-track is necessary to save the five men on the main track, because the side-track immediately re-unites with the main track after the point where the lonely victim is tied and before the point where the five are. So, if there were no lonely man tied to the side-track, turning the trolley using the switch would have no effect to save the five from death. Contrary to what the DDE suggests, people do not condemn turning the trolley in *Loop* as much as they do condemn pushing the bystander in *Footbridge*. The sample is usually divided fifty-fifty (roughly) when participants are asked to assess an interventionist option on *Loop*.

of judgments by emotions. To my knowledge, Mikhail never replied, but I suppose he can reply in no convincing way. So far, there is no theory that fits the 'trolleyology' data better than Greene's.

Secondly, Dupoux and Jacob say that moral dumbfounding is good evidence for both Haidt's SIM and UMG, so that that kind of evidence cannot adjudicate between the two theories. I wholeheartedly agree with them on this point. Furthermore, they implicitly follow the lines of Pizarro, Bloom, and Fine in their criticisms to the SIM, arguing that there are cases in which explicit moral principles do matter in moral judgments. According to Dupoux and Jacob, this creates a disanalogy between Chomskian grammar and UMG. Chomskian grammar posits a strong distinction between deep, implicit grammatical structures and folk-ethnological grammatical principles, but we cannot find this distinction in moral cognition. Another strong disanalogy between UMG and transformational-generative grammar is given by intractable moral dilemmas. In Dupoux and Jacob's opinion, all linguistic dilemmas are solved in a pragmatic way, whereas this is not possible for moral dilemmas. The presence of moral dilemmas indicates that the moral module does not work as a linguistic module *à la* Chomsky. A third point is that moral cognition violates various requirements for Fodorian modularity. Moral judgments violate encapsulation, because any added background information about an action or an omission may alter moral evaluation. It also violates compositionality, because the deontic value of an action cannot be deduced from the deontic value of its components. The single components (for instance, to use Dupoux's and Jacob's example, brewing tea, opening a bottle, pouring the content in a cup of tea, serving the tea to a guest) may have *per se* moral valences that are very different from the moral valence of the whole action (poisoning a guest to death). Because of all of these disanalogies, the idea of a UMG does not seem particularly helpful to explain moral cognition. Dwyer and Hauser (2008) try to defend themselves by saying that they do not necessarily subscribe to the idea that the moral module is a Fodorian module. Nonetheless, their reply does not defuse the other

disanalogies spotted by the French cognitive scientists, nor does it do anything to avoid the clash with 'trolleyology' and with the experiments showing emotional modulation of moral judgments.

As to Sterelny (2010), he claims that morality is an adaptation, but this does not involve innateness. Adaptations can also require a big amount of learning to work properly. Sterelny admits that morality is an adaptation, but denies that it is innate. In his view, moral judgments are the result of generalizations from learned exemplars through pattern recognition. The child is exposed, e.g. through tales, to moral exemplars and then learns to recognize corresponding paradigmatic patterns of behavior in everyday life. In this way she joins a moral community, a community that has engineered her learning environment so that she and her fellow kids learn the values the parents already share. Sterelny stresses that moral judgments have gradients and are not categorical, unlike grammaticality judgments. This criticism seems to be warranted and adds up to the list of disanalogies already highlighted by the work of Dupoux and Jacob. Sterelny does not provide a spot-on criticism of the UMG, as Greene and the French cognitive scientists do. On the contrary, he offers an alternative view, a view that is alternative also to Haidt's MFT, as both UMG and MFT are nativist. I find Sterelny's position interesting because his work shows how underdetermined theories in experimental moral psychology are. We still need many data points to narrow the number of plausible theories down to a manageable amount. At the moment, very different theories still have a right of citizenship in the debate.

Summing up, I think that, given the amount of disanalogies between grammatical capacity and moral capacity, the lack of compelling empirical evidence in favor of UMG[86], and serious problems to fit with experimental results, UMG ought to be considered as an interesting working hypothesis which is not particularly substantiated at the moment.

In closing this chapter, I would like to briefly deal with the position expressed by neuroscientist Michael Gazzaniga (2005). Gazzaniga draws on Hauser and claims that most

[86] I.e. empirical evidence that only UMG can explain.

ethical systems across history and around the globe evaluate certain things in similar ways. Almost the totality of human groups have held murder and incest as morally wrong, have condemned lack of care towards children, have chastised lies and violations of trust. Gazzaniga stresses the importance of the controversial concept of human nature and underlines similarities among specimens of *H sapiens*. He (2005, xix) posits that there is

> a universal set of biological responses to moral dilemmas, a sort of ethics, built into our brains. My hope is that we soon may be able to uncover those ethics, identify them, and to begin to live more fully by them. I believe we live by them largely unconsciously now, but that a lot of suffering, war, and conflict could be eliminated if we could agree to live by them more consciously.

In particular, humans usually live by these ethical principles unconsciously, but if humans spelled them out more clearly through empirical investigation, we could live by them explicitly, in a conscious, reflective way.

In this last part of this section, I examine Gazzaniga's neuromoral theory. His idea is that the implicit principles of UMG sometimes fail to influence behavior. This brings about "a lot of suffering, war, and conflict." What could we do to avoid this? Quite simply, apply the principles of the UMG in a conscious way. Hence, explicit knowledge of the principles (acquired through psychological and neuroscientific investigations) could allow us to make better moral judgments, i.e. moral judgments that lead less often than at present to "a lot of suffering, war, and conflict."

A first question that might be asked is whether it is *de facto* possible to apply those principles in an explicit way. It might be the case that human conscious computational power is too little to perform explicitly the processing required by implicit UMG principles. We experimentally know that the computational power of conscious reasoning is close to zero if compared to the computational power of implicit areas of the brain such as V1, the primary visual cortex in the occipital lobe (Dijksterhuis and Nordgren 2006, 96-

97). Of course, it might also be the case that the opposite is true and that it is very well possible to follow those principles in a deliberate way. This is an empirical question and only experimentation can eventually address it. However, it cannot be taken for granted that what Gazzaniga proposes is feasible.

Secondly, even assuming *arguendo* that what Gazzaniga suggests to do (i.e. to move from the implicit mode of making moral judgments to the explicit mode of making moral judgments) is possible, his theory runs into what I dub as the 'meta-normativity problem'. This is a recurring problem of neuromoral theories. Gazzaniga's suggestion amounts to this: "We *ought* to make moral judgments by deploying the principles of UMG in an explicit way rather than by deploying them in an implicit way." What kind of normativity does this 'ought' express? There are at least two possibilities.

The first possibility is that this claim is in turn a moral judgment. Yet this paves the way to new questions that a neuromoral theorist of this sort ought to answer: is this moral judgment true? And what resources can this neuromoral theorist use to justify it? Suppose a theorist of this kind suggests that our grasp of the normative moral standards derives from UMG itself. Would this mean that the only way we have to justify the moral claim "we ought to consciously apply UMG's principles" is by appeal to the principles of UMG? If so, would this make the justification problematically and viciously circular? Or would the circularity be acceptable? I do not want to answer these questions here. It is up to this sort of neuromoral theorist to answer them.

Another possible option is that, when this kind of neuromoral theorist says that one ought to consciously deploy the principles of UMG, the 'ought' in question is not intended to be a moral one but rather a prudential[87] one. The claim would be that the deliberate

[87] With 'prudence' I do not indicate a normativity that aims at self-interest, but one that aims at maximizing desire satisfaction. There are many desires that people have that are about the well-being of other people. So one may desire that one's relatives or friends are successful and also desire this in itself, and not because their success will bring one benefits. One may also desire that people in a far away country - whom one will never meet and whose life will have no impact on one's own life - live good lives and are not, for example, killed by disease, famine, or war. Such desires are altruistic, prosocial. Still, there are things that are prudentially good in relation to them - things that lead to their satisfaction - and things that are not. So, under my reading of prudence, there is no conflict between prudence and pro-social attitudes. My reading is reminiscent of the

deployment of the principles of UMG is the best way people have – either as individuals or as collectives – to secure some desired outcomes, such as the reduction of suffering and the avoidance of conflict in the world. Given the phrasing of Gazzaniga's claim, this is likely to be the correct interpretation of his position. But if this is the right way of interpreting the 'ought', there are other questions one can ask. In particular, one can ask whether the claim is actually true. Is it actually true that deploying those principles in a conscious way would lead to *better* moral judgments, where 'better' refers to prudential standards? Let us grant, for the sake of argument, that there is such a thing as a UMG. What are the reasons for thinking that UMG does not contain principles that, say, in contemporary societies are likely to lead, at least in some circumstances, to conflict, suffering and other unpleasant things that people – either individually or collectively – would prefer to avoid? This is an empirical claim, and besides an empirical claim that is very difficult to test.

Gazzaniga's neuromoral theory incurs into both an empirical problem of feasibility ('can we actually deploy the implicit principles of the UMG in an explicit way?') and the conceptual 'meta-normativity problem'. Gazzaniga seems to give a prudential answer to the meta-normativity problem, but this in turn opens up the big empirical question about the effectiveness of conscious application of UMG as a tool to pursue prudential goods. So Gazzaniga faces now two important empirical questions and his neuromoral theory has to address them if progress is to be made.

If Hauser, Mikhail, Dwyer, and Gazzaniga take morality as composed of grammatical rules or implicit moral principles, the theorists I examine in the next section focus on fast and frugal procedures called heuristics.

3.4. Moral heuristics: friends or foes?

---

one put forward by Joyce (2001). For a critique of the Kantian view of morality as a set of categorical imperatives and an attempt to build a morality on hypothetical imperatives, see Foot (1972). Foot's idea of morality shares some similarities with the idea of prudence I am describing in this footnote.

In experimental psychology the word "heuristics" usually appears in the "heuristics and biases approach" championed by two of the most influential psychologists of the last century, the late Amos Tversky and Nobel laureate Daniel Kahneman. The heuristic and biases approach, based on a series of famous studies published in the 1970s and in the 1980s (e.g. Kahneman and Tversky 1979; Tversky and Kahneman 1974, 1981, 1983), aims at studying human rationality by means of the errors it commits. Kahneman and Tversky found that people behave irrationally in many circumstances: they are often insensitive to sample size and prior probabilities in statistical judgments, they do not follow basic logical rules such as the conjunction rule[88], and are influenced by the wording of outcomes, since they prefer to avoid losses than to forfeit gains, if their absolute magnitude is the same. Some of these results, especially those concerning the different attitude of humans towards losses and gains and the fact that humans are risk-takers when they want to avoid risks and risk-averse when gains are involved[89], are explained by prospect theory (Kahneman and Tversky 1979). Some other results are explained through heuristics, i.e. fast decision making procedures that take into account just a tiny part of the information potentially available to the agent and regularly use just fragments of incomplete information. The best known heuristics are the representativeness heuristic, according to which one item is associated with a category if it looks like the stereotype for that category[90], and the availability heuristics, according to which things that more easily come to mind[91] are seen

[88] According to which "A&B" needs to be *less* probable than "A" and *less* probable than "B".

[89] One example is the "Asian flu" case I mentioned above in Ch. 2. If outcomes are described in terms of "deaths", they are seen as losses and people take risks; if they are described in terms of "lives saved", they are seen as gains and people become risk-averse.

[90] E.g. participants are given a description of a shy, meticulous man, are asked whether he is a librarian or a farmer, and are told that this man comes from a group in which there are more farmers than librarians. The man is regularly judged as a librarian by participants even though there are much more farmers than librarians in the population at large and the experimental participant knows this (cf. Tversky and Kahneman 1974).

[91] For instance, in English there are more words whose third letter is 'r' than words whose first letter is 'r'. However, participants find the latter much easier to remember and hence usually think that there are more words which start with an 'r' than words whose third letter is an 'r'. (cf. Tversky and Kahneman 1974).

as more frequent than things that less easily come to mind. There are of course many other heuristics, but this can be enough as an example. The final outcome of this approach is deeming most human beings irrational, relative to a standard that is constituted by RCT, the standard theory of rationality in games theory and micro-economics since the end of WW2. According to RCT, humans should be insensible to wording effects and just maximize their individual utility. Hence, Kahneman and Tversky see heuristics as deviations from the norm, i.e. as something bad relative to the ideal represented by the rationality of the sheer maximizer. This claim of widespread irrationality has given rise, mostly in the 1990s, to what in psychology are known as the 'rationality wars' (Krueger and Funder 2004; Stanovich and West 2000; Stein 1996). The front opposing Kahneman, Tversky, and most economists in the world features German psychologist Gerd Gigerenzer and his numerous followers. Gigerenzer heavily draws on the work of the late Nobel laureate Herbert Simon (1955, 1956), who famously criticized maximizing rationality by creating the concepts of 'bounded rationality' and 'satisficing'. Satisficing is picking the first option that the decision-maker encounters which in turn satisfies some conditions she posited, i.e. choosing the first option she finds that is 'good enough' for her. Simon's main take is that humans do not possess sufficient cognitive resources to carry out maximization and that maximization would be largely unnecessary for survival of organisms such as *H sapiens* specimens in the environment. Gigerenzer follows him on this path and claims that heuristics are not at all deviations from rationality. To quote the title of Gigerenzer et al. (1999), heuristics are what makes us smart. According to Gigerenzer the human mind can do little better than using heuristics. Heuristics make our survival in the environment possible and of course they bring about mistakes, but these mistakes are inevitable for creatures such as humans, so that we ought not to weep on them (cf. Gigerenzer 2005).

Heuristics are highly context-sensitive, and together with a description of the environment in which the choice takes place, they can often explain an agent's behavior. They are more conducive to success of an organism in the environment, defined as a match between behavior and environment enabling survival and eventually reproduction, than the informationally rich processes of conscious deliberation (cf. Gigerenzer 2000; Gigerenzer and Selten 2001). Thus heuristics are very rational, not irrational as Kahneman and Tversky have been claiming, and the RCT normative standard needs to be jettisoned.

The discussion about moral heuristics is set in this context and features as main characters Gigerenzer himself and American legal scholar Cass Sunstein (2005), who takes the side of the 'Kahnemanian' mainstream. Gigerenzer (2008a, 2010) claims that most of humans' decisions in the moral domain are driven by heuristics, exactly as it happens, according to his view, in most domains of human cognition. Heuristics ought not to be confused with Haidtian moral intuitions. Gigerenzer is rather skeptical about the concept of moral intuition and about dual-process talk in general (2008a, 15). He claims that moral intuition in the Haidtian sense is not a primitive notion, as Haidt takes it to be, and that intuitions are actually the result of the covert operations of moral heuristics. The heuristics people deploy to solve moral problems and to form moral judgments are not specific for thinking only about moral issues, so that there are no "moral heuristics" *sensu stricto*. Moral problems get solved by the same heuristics that are used to solve other kinds of social problems. So, in this perspective, heuristics make us smart and, at the same time, make us moral. However, Gigerenzer (2008a, 3) claims that the very same heuristics can in different circumstances result in morally good or morally bad decisions. For example, a simple rule such as "if there is a default option, do nothing about it (i.e. tacitly accept the default)" may result in morally good or morally bad decisions, depending on what the default option

actually is. In one of Gigerenzer's examples, an opt-out policy for organ donation[92] will result in most people's tacit acceptance of the default option, that is, a (perhaps unconscious) decision in favor of organ donation. The same applies to heuristics such as "do what the majority of the members of your group do". In a society in which most people behave in racist ways, this heuristic will result in racist behavior. As to the domain of morality, Gigerenzer claims that what counts as morally relevant or irrelevant is highly variable. There is a mobile 'moral rim' that separates issues of taste from moral issues. This rim moves according to historical periods and cultural groups. This is due to the fact that there are no specific moral heuristics. Gigerenzer claims that the rationality standard he contests in his debate with Kahneman and Tversky (Gigerenzer 1991, 1996; Kahneman and Tversky 1996) is used both in non-moral normativity (i.e. as a norm of what counts as rational) and in moral normativity (i.e. as a norm of what counts as morally mandatory). Thus, Gigerenzer makes two claims: (1) psychological theories that assume that a human agent *is* a utility maximizer are descriptively inadequate; (2) normative moral theories that maintain that a human agent *ought to be* a utility maximizer are normatively inadequate. As we will see, claim (2) amounts to a neuromoral theory. These two claims are closely connected, as in Gigerenzer's opinion there is a strong genealogical link between RCT and consequentialist moral theories. RCT genealogically stemmed from normative (and not descriptive) frameworks that were popular in the Enlightenment Age, such as, indeed, Bentham's maximizing consequentialism. So, these descriptive and normative views can be considered as parts of a single intellectual stream. Hence, RCT and similar descriptive theories of human choice conflate a *desideratum* for actual reality, giving rise to an

[92] In other words, a policy according to which all citizens are by default potential organ donors.

egregious case of wishful thinking, or, as Gigerenzer puts it, to an instance of "ought-to-is transfer" (2010, 532).

Gigerenzer deems utility maximization impossible for humans. These are the reasons he provides. First, calculating the expected utility of different possible actions is in general a computationally intractable problem, as the possible outcomes are too many and probabilities hard to figure out. Secondly, utility has various components, such as health, income, freedoms, opportunities, and so on. Gigerenzer claims that, given that such components are in conflict with each other (i.e. there are trade-offs among them) and incommensurable (i.e. there is no common currency that can be used to weigh one component against the others), the maximization of utility will in general be impossible. As I have written above, maximizing utility is (in Gigerenzer's opinion) possible only in "small worlds" (cf. Gigerenzer 2010), i.e. artificial scenarios such as trolley dilemmas and economic games, where payouts of actions and probabilities thereof are known, whereas in "large worlds", such as the real one, the decision-maker lives in a constant lack of relevant information about outcomes value and outcomes probability and tries to survive by availing herself of the little information and computational power she can marshal. Furthermore, Gigerenzer argues that maximization is not even a good ideal benchmark for what constitutes a good decision: not only it is impossible to compute, but one ought not even to try to approximate it. Gigerenzer quotes here Lipsey's (1956) general theory of the second best. In a situation in which fulfilling all the conditions required to achieve the optimal outcome is not possible, trying to fulfill as many as possible of such conditions is not necessarily the second best option. The second best option has usually features that are very different from the ones of the first best option, so that trying to realize as many traits as possible of the first best option does not help one to obtain the second best result. In

other words, doing things that in ideal conditions would lead to maximization is not a good way of trying to achieve our aims. Then Gigerenzer also argues that, in some cases in which maximization of monetary outcomes is feasible because the variables involved come in a manageable number, satisficing may outperform maximization. Satisficing strategies and heuristics regularly suffer for *bias*, i.e. they introduce systematic errors that are always in the same direction. In contrast, attempts at maximization suffer from *variance*, i.e. error in estimating the parameters of the population of interest from the sample data. Hence, in these cases, there is a bias-variance dilemma, and in a subset of these cases it may be better to have bias than variance.

Let us now move to Gigerenzer's neuromoral theory. One way in which knowing the machinery for moral judgments can help people make good moral judgments is the following. A proper understanding of the mind and of its heuristics 'toolbox' will help one see that many of the current theories of decision making – including many theories of *moral* decision making – are wrong. They are wrong both in that they do not accurately represent the way people make moral decisions ("descriptively") and in that they give the wrong advice about the proper way of making decisions ("prescriptively"). An understanding of the heuristics toolbox will help one realize that one *ought not* to follow the advice of such theories when trying to solve moral decision making problems. According to Gigerenzer, one ought to rely on the heuristics toolbox instead. Trying, via conscious deliberation, to counteract or bypass the outputs generated by the heuristics toolbox is a bad idea. In arguing for this, Gigerenzer (2008a) discusses maximizing act utilitarianism. He takes this view to imply that in order to decide what ought morally to do, one has to determine the expected aggregate utility of all the available courses of action and then choose the one with highest value. But this is impossible for the reasons I have

explained above. So a proper understanding of the machinery for moral judgments may help us not be led astray by bad theories such as maximizing act utilitarianism. But there is also another, and more positive, way in which an understanding of the heuristics toolbox can have a positive impact on our moral-judgment-forming ability. If heuristics can give rise to good and bad judgments depending on the context, then arguably knowledge of the toolbox could be used to maximize the chances that people will find themselves in contexts in which the heuristics they will tend to use will result in good judgments. Knowledge of the heuristics toolbox can be used in this way both by individual agents and by policy makers who have some control on the contexts in which certain populations make choices (cf. the notion of 'choice architect' in Sunstein and Thaler 2003). That is, if a person can affect the context in which decision making takes place, either for oneself or for others, he can use an understanding of the heuristics toolbox to affect such context and, specifically, to affect it in a way that is conducive to good moral judgments (cf. Gigerenzer 2010; Todd and Gigerenzer 2007). The focus here is on shaping the decision making *context* rather than on changing the heuristics themselves, which in Gigerenzer's view are triggered automatically, in the sense that their deployment comes natural and instinctive to human beings and – in the absence of interference at least – may be difficult to avoid. This being said, Gigerenzer's neuromoral theory raises some questions that are similar to the ones evoked by Gazzaniga's view. The meta-normativity problem applies again: what is the normative standard by which one should evaluate the maximizing act utilitarian's recommendations?

Before moving forward, it must be noticed is that maximizing utilitarianism need not be conceived as a theory of *how one should determine* what to do, as a theory in moral epistemology. It can also be conceived as a theory of *what is* the morally correct thing to

do, independently of how one determines that it is, as a metaphysical theory. This metaphysical variety of maximizing utilitarianism is an account of the nature of the normative standard that applies to actions and decisions. On this interpretation, the maximum can be used as a benchmark to assess the goodness of choices, even though it cannot actually be computed in everyday decision making.

On one version of metaphysical maximizing utilitarianism so conceived, an action is morally right if it maximizes aggregate utility. Since trying to make evaluations and comparisons that are too difficult for creatures like us is certainly *not* conducive to the maximization of utility, a maximizing utilitarian of this sort can use Gigerenzer's own arguments and recommend not doing all the calculations that Gigerenzer also recommends not doing. In particular cases, this utilitarian can also recommend using the heuristics that Gigerenzer describes, at least if some degree of control on the heuristics one deploys is possible – whether it is direct or indirect control, via environmental manipulations. This kind of theorist can argue that one should use the heuristics if and when doing so is the action that maximizes aggregate utility.

On another version of metaphysical maximizing utilitarianism, an action is right if it is produced by decision making procedures whose deployment – in organisms like us and in a world like ours – tends, at least in the long term, to maximize aggregate utility. Again, this kind of utilitarianism is compatible with Gigerenzer's focus on heuristics. If the heuristics turn out to be the decision making procedures whose deployment tends to maximize aggregate utility, then a utilitarian of this sort can recommend using them[93].

[93] The distinction between the two versions of metaphysical maximizing utilitarianism is akin to the distinction between act-utilitarianism and rule-utilitarianism, at least on *some* ways of understanding the latter distinction. Both of these ways of interpreting utilitarianism exclude that maximization should be a procedure for making decisions. On the contrary, the utility maximum is just taken to be a benchmark. On the first interpretation, what gets measured against the benchmark are actions, single instances of behavior. This is vaguely similar to act consequentialism: an action is morally required if its outcome gets closer to realizing the utility maximum than competing alternatives. On the second interpretation, what gets measured against the benchmark are decision making procedures. This is roughly reminiscent of rule consequentialism: the

Gigerenzer makes the following comment on the view that maximization can be seen as a normative standard (as opposed to a *procedure* by which to determine whether one should perform particular actions or adopt particular decision making procedures):

> I must admit that I fail to understand the logic. […] Even if someone were to stumble over the best action by accident, we would not recognize it as such and be able to prove that it is indeed the best. How can maximization serve as a norm for rightness if we can neither determine nor, after the fact, recognize the best action? (Gigerenzer 2008b, 44)

If Gigerenzer is right about maximization being computationally out of reach for creatures like us and in a world like ours, then metaphysical maximizing utilitarianism can be accepted only by accepting the claim that humans can never have knowledge of moral facts, because – due to their computational limitations – they can never have knowledge of what action or what decision making procedure is the one that maximizes utility. This seems to be a hard bullet for the metaphysical maximizing utilitarian to bite, and so such a utilitarian would presumably try to argue that maximization is *not* computationally out of reach for organisms like us, at least not in general and systematically. Independently of human epistemic and computational limitations, Gigerenzer also suggests that, at least in some cases and for reasons that have to do mainly with incommensurability, maximization is impossible, in the sense that *there is no fact of the matter* as to which action (or which decision making procedure) would maximize utility. If he is right on this, then the metaphysical maximizing utilitarian has to accept that – at least in some cases – maximization *cannot* constitute a normative standard. A metaphysical utilitarian could of

---

adoption of a decision making procedure is morally required if the outcome, in the long run, of that decision making procedure gets closer to realizing the utility maximum than competing alternatives. Of course, this leaves open the possibility that adopting a given heuristic is morally required.

course deny Gigerenzer's claims on incommensurability, or he could give up on maximization *in specific cases*, those affected by incommensurability problems, as long as he can argue that such cases are relatively uncommon. I do not want to enter this debate. I concede that Gigerenzer's claims, when properly reconstructed, *can* be seen as a critique of the maximizing act utilitarian's account (or accounts) of the normative standard for actions and decision making procedures. But, if so, an alternative account of the nature of such a normative standard needs to be provided.

This being said, let us go back to the 'meta-normativity problem'.

One option is that Gigerenzer's meta-normativity is in turn moral. If this were so, Gigerenzer's normative claims would express moral judgments. Then one could ask whether these moral judgments are correct, what resources could be used to justify them, and whether one could appeal to the heuristics 'toolbox' in this context without incurring in circularity. Unfortunately, Gigerenzer does not explore these issues.

Another option is that Gigerenzer's meta-normativity is not moral, but prudential instead. The view being proposed then would be that relying on the heuristics toolbox and designing choice environments in certain fashions is the best way people have – either as individuals or collectives – to secure certain important desired outcomes, such as peaceful relations, stable and rewarding cooperation, etc. Is this Gigerenzer's favorite option? He refers to what he calls "ecological rationality." Heuristics are ecologically rational in the sense that, in most environments, they are the best or the only way to achieve "success." As I have written above, success is about a match between a given organism and the environment; furthermore, it is measured in terms of "accuracy, frugality and speed of decisions" and involves "means suited to certain goals" (Gigerenzer and Sturm 2011, 13). But what are these goals? Gigerenzer suggests that some important goals for judging the

ecological rationality of heuristics are desire satisfaction and the fixation of true beliefs (*ibidem*). He seems to be claiming that we ought to rely on the heuristics toolbox in the sense that doing so is the best way we have to satisfy our individual or collective desires, including those desires that concern or involve attaining true beliefs.

In an article on 'moral satisficing', Gigerenzer claims:

> My aim is not to provide a normative theory that tells us how we ought to behave, but a descriptive theory with prescriptive consequences, such as how to design environments that help people to reach their own goals. (Gigerenzer 2010, 530)

This statement is somewhat puzzling, in that a prescriptive theory is normally understood as one that prescribes actions and tells people what they ought to do. I believe that Gigerenzer is trying to say something like the following. People can, either individually or collectively, select certain goals, acquire certain desires. Some of these goals are labeled "moral goals" and involve certain kinds of behaviors and outcomes. The labeling in Gigerenzer's view is not very important: it is just a way of marking certain desires as particularly significant and different cultures do it in different ways (Gigerenzer 2010, 543). Once a moral goal has been selected, one has to identify the best means of realizing it. The distinction Gigerenzer makes between normative and prescriptive is the distinction between a theory that tells you what you ought to do when you are trying to select your moral goals and a theory that tells you what you ought to do in order to realize a moral goal that you have already selected. He only wants to give a theory about what you ought to do in order to realize moral goals that you have already selected. Having clarified this, it must be noticed though that, once Gigerenzer's meta-normative standards are interpreted as prudential, some important issues come into focus. Gigerenzer seems to be saying that no matter what individual or collective goal is selected, simple heuristics are going to

outperform informationally rich strategies and conscious reflection. But his arguments against maximization do not support this strong and general claim. In order to see why, I draw on some criticisms Cass Sunstein makes on Gigerenzer's view.

Sunstein (2005, 2008) objects that certain goals – especially collective goals that are determined through complex negotiation and collective deliberation – seem to be such that humans are more likely to realize them through slow and informationally rich conscious reflection. In particular, I add, conscious reflection seems to be the best tool social engineers and choice architects have for affecting environments in ways that will lead humans to decide and act – perhaps via the deployment of simple heuristics – in ways that are in turn conducive to outcomes that are desirable under a prudential standard. In other words, even if I granted *arguendo* that Gigerenzer is right at the level of everyday behavior and choice, this would not mean that he is also right at the level of choices about the ways in which environments need to be shaped. For instance, suppose that some public health officers must decide how to shape the decision environment for people that must choose between giving their consent to organs being explanted from their corpses and not giving such consent. It may be the case that conscious deliberation guides the officers' decision better than a heuristic, since they have to take into account humans' psychology, different trade-offs between individual autonomy and social beneficence, the interests of relatives, and so on and so forth. At least, this is what Sunstein is arguing for. In his opinion, in order to evaluate and compare heuristics and the outcomes they generate in different possible choice environments, policy makers may well have to perform informationally rich analyses, some of which could appeal to Gigerenzer's own research. A neuromoral theorist *à la* Gigerenzer needs to make these distinctions and to assess their normative implications. Moreover, Sunstein makes another point. In accord with Baron (1994), he sees heuristics as undue generalizations. Some decision making procedures work well in most cases, and are hence generalized to all cases. However, this extension can prove dreadful in exceptional cases and bring about very bad consequences. For instance, "one ought not to

lie" works very well in many cases, but when an armed man who wants to kill your best friend asks you where she is, it works less well. Sunstein claims that moral heuristics could lead to moral errors and these errors ought to be avoided. As we have seen, Gigerenzer claims that these errors are a necessary part of the heuristic toolbox: if humans are to be intelligent, they also need make mistakes. Sunstein has a more traditional view on rationality, and hence thinks that moral mistakes are at least partly avoidable. Yet one must ask what it means for a heuristic to work 'well' or 'less well', or, which is equivalent, what is the meta-normative standard that defines what a moral error is. Sunstein recognizes that there is strong disagreement among both psychologists and philosophers about what constitutes a moral error. He claims, this time contrary to Baron (1994), that it is not possible to use a controversial standard such as straightforward consequentialism to assess what amounts to moral error. I agree with him on this. So Sunstein tries to locate a non-controversial meta-normative standard to assess the goodness of heuristics. I am pleased by the effort and less pleased by the results. Sunstein puts forward a notion of 'weak consequentialism' according to which the consequences of a decision making procedure do matter and violations of rights and duties can be factored in the decision making process as consequences. This standard may admittedly reach a very high degree of consent, but it seems to beg the question against a staunch nonconsequentialist, such as Kant (1966/1797)[94]. Hence, weak consequentialism does not seem to help us much address the meta-normative issue. I find more helpful a suggestion by Pizarro and Uhlman (2005): if an agent A understands that her moral judgment depends on factors that A herself considers to be irrational or irrelevant, the judgment can be said to be in error. This is a plausible, albeit subjective, view on moral error. It will capture at least some cases of moral error.

[94] Another kind of nonconsequentialist, whom I could label as the 'mild nonconsequentialist', may not be committed to the claim that the duty not to lie is absolute – it could be a *prima facie* duty that might be trumped by your friend's right to life and by special obligations you have implicitly stipulated with her due to your friendship.

Summing up, Gigerenzer is likely to support a prudential meta-normative standard, but does not address the very important problems this position brings forth, in particular the empirical issue that conscious deliberation is not conducive to the pursuit and obtainment of prudential goods such as peaceful relationships and stable cooperation. On the other hand, Sunstein's attempt at finding a non-controversial meta-normative standard for decision making procedures is highly praiseworthy, but the attempted solution, weak consequentialism, does not live up to expectations because it begs the question against staunch deontologists, who cannot simply be dismissed as irrational.

In the next section I describe views that, as Gigerenzer's, attribute importance to unconscious thinking.

3.5. Unconscious Thought: Dijksterhuis, Woodward, and Allman

Unconscious Thought Theory is championed by Dutch psychologist Ap Dijksterhuis (Dijksterhuis and Nordgren 2006; Dijksterhuis et al. 2006). It is a theory in general experimental psychology and not only (and not necessarily) in experimental moral psychology. The theory posits that humans are capable of conscious and unconscious thought. Unconscious Thought (henceforth UT) can be defined as thought or deliberation in the absence of conscious attention directed at the problem. In contrast, conscious thought is defined by thought in which attention is devoted to the object or the task. In other words, conscious thought corresponds to cognitive or affective thought processes that occur while the object or task is the focus of one's conscious attention. UT is what, in popular wisdom, is labeled as "sleeping on a problem." Dijksterhuis claims that UT can, in a variety of circumstances, address problems better than conscious thought. Surprisingly, UT is better at dealing with complex problems, where lots of attributes for each available option are in play, than with simple problems, where conscious reasoning performs better. This happens because consciousness has a low capacity: it cannot store much information. In contrast, UT does not suffer from low capacity, so that the quality of choices does not

deteriorate with increased complexity. The theory is backed by experimental results. In Dijksterhuis et al. (2006), participants were either in the conscious thought condition or in the UT condition. Participants who were assigned to the latter condition were asked to choose an option among a list of cars and were allowed some time to reflect, but were distracted by a cognitive load task (for instance, solving anagrams) for the whole duration of the reflection period. So the participants in this condition were not able to reflect at all. The cars could be characterized by either a long or a short list of attributes. Conscious thought is better at selecting among cars that are described through a few attributes each than at selecting among cars which are coupled with a long list of attributes. Another feature of conscious thought is that, due to its poor capacity, it induces biases: it leads people to put disproportionate weight on attributes that are accessible, plausible, and easy to verbalize. In other words, conscious thought is more vulnerable than UT to 'Kahnemanian' biases. However, there are some tasks that only conscious thought can solve, e.g. arithmetic tasks. UT cannot follow the strict rules that are required to perform these computations, whereas conscious thought is very good at that. On the contrary, UT excels in pattern recognition, which does not require precise rules to be followed.

All of this seems to bring grist to Gigerenzer's mill, but it is not actually so, since Dijksterhuis and Nordgren (2006, 104) specify that heuristics are carried out by conscious thought and not by UT. So the results by Dijksterhuis do not necessarily support Gigerenzer in his claims against maximizing act consequentialism, since both contenders (Gigerenzer and the maximizing consequentialist) are on the side of conscious thought in the conscious vs. unconscious thought dichotomy.

Nonetheless, UT is connected with Haidtian moral intuitions, in the sense that these intuitions are the result of UT. This idea is further elaborated by Woodward and Allman, as I explain below. At variance with that, Dijksterhuis and Nordgren propose a normative view concerning general cognition: "One should give more weight to the unconscious intuitive feeling than to the conscious pluses and minuses." (2006, 107).

For the purposes of this dissertation, the interesting point about Dijksterhuis's views is the way in which it challenges Greene's dual-process model. In particular, it characterizes two systems of thought that do not fit particularly well into Greene's scheme. System1 is fast and inflexible, but UT seems to be at least as flexible as conscious thought, for example, and not to be necessarily fast, since UT yields better results if it is allowed more time to process information, i.e. if 'incubation' takes longer (Dijksterhuis and Nordgren 2006, 99). So, if Dijksterhuis's general theory is valid, Greene's descriptive model is undermined. Yet it is not clear whether Dijksterhuis is right. Lassiter and colleagues (2009) carried out two experiments to test an alternative explanation of UT. UT would consist, according to their hypothesis, in just recalling an on-line decision that has been made when the stimuli were presented to the participants. All participants in the original experiments by Dijksterhuis and co-workers were allowed to look at the experimental materials, i.e. the data about the cars, for a short time only, and at the same time asked to form a first impression. Then participants in the conscious condition had 4 minutes to "think carefully," but they could access the material no more. However, they had not memorized the data, so that they were actually unable to carry out a conscious decision making process. They could not memorize because they were asked by the experimenters to form a first impression (as opposed to memorize) in the time when the experimental material was accessible. So their decision making process was impaired by the experimental setting Dijksterhuis and colleagues created. The good performance of the UT thinkers is, in Lassiter's and co-workers' opinion, just an artifact due to a poor experimental design. Strick, Dijksterhuis, and Van Beeren (2010) replied with other experiments. Deciding who is right between the Ohio guys led by Lassiter and Dijksterhuis's lab in Nijmegen (The Netherlands) would require delving deep into experimental details, and this, however interesting, would lead us too far away in this context. However, it is important to notice that the main experiment by Dijksterhuis and colleagues (2006) lacks an important control: there is no 'immediate judgment' condition in which participants were asked to make a selection forthwith. In

spite of this important shortcoming, the UT issue is still open because Nordgren, Bos, and Dijksterhuis (2011) have recently found that a combination of conscious and unconscious thinking is even more efficacious than UT alone at solving complex problems in which lots of attributes are involved. However it may be, this line of research suggests the possibility that UT is likely to be more than a series of automatic responses, as Greene instead posits.

Let us move now to the neuromoral theories that have been created starting from these descriptive (and non-morally normative) views. James Woodward and John Allman (Allman and Woodward 2008; Woodward and Allman 2007) mainly deal with moral intuitions, that are defined along Haidt's (2001) lines. As in Gigerenzer's view, moral intuitions are for them a part of a broader set of social intuitions; moral cognition counts as a sub-set of social cognition. Moral intuitions are deeply intertwined with emotions and are created by a cognitive system that can process a "large number of disparate social cues" at high speed (Woodward and Allman 2007, 182). Moral intuitions are the result of implicit learning, so that, at variance with Gigerenzer's social and moral heuristics, they are not necessarily frugal and depend on exposure to social stimuli to a greater extent than Gigerenzer's heuristics seem to do[95]. Moral intuitions arise through a process that is remarkably similar to Dijksterhuis's UT, even though Woodward's and Allman's interpretation features emotions in a prominent role. On the contrary, UT is not necessarily emotionally-loaded in Dijksterhuis's descriptive hypothesis. This difference notwithstanding, Woodward and Allman *de facto* extend Dijksterhuis's ideas to the moral case. They believe that moral intuitions relying on UT can lead to better decision making processes than more conscious procedures, if the amount of relevant information to be taken into account is massive and implicit learning is possible. In turn, such learning is possible when an individual has been, directly or indirectly, in contact with the relevant kind of experience. Given the limited capacity of conscious thought and the complexity of most morally relevant situations, conscious reasoning has little hope to come up with

[95] Yet Gigerenzer is never, at least to my knowledge, very clear about the development of heuristics and about how robustly they appear in children.

satisfactory decisions concerning these cases. At variance with that, moral intuitions can process in parallel big quantities of environmental and social cues, so that they usually result, according to Woodward's and Allman's view, in better moral decisions than explicit cost-benefit analysis.

This amounts to a neuromoral theory: deeper knowledge of the machinery for moral judgments (i.e. knowing what cognitive processes moral intuitions are based on) leads to better moral judgments. Various considerations underpin this neuromoral theory.

First, a positive aspect of moral intuitions is that they can take into account effects of our decisions on the future behavior of other human beings that, in Woodward's and Allman's opinion, would be difficult to figure out by making use of conscious cost-benefit analysis. For instance, negotiating with terrorists or paying a ransom to a kidnapper now might encourage further terrorist acts and further instances of kidnapping in the future. This might be missed by explicit cost-benefit analysis, but moral intuitions are likely to be perceptive to this aspect, albeit in an unconscious way.

Secondly, Woodward and Allman distinguish two versions of consequentialism: strategic consequentialism and parametric consequentialism[96]. Strategic consequentialists understand that moral contexts imply an "ongoing interaction with other actors who will respond in complex ways that are not easy to predict, depending on the decision maker's choices […] these responses will in turn present the original decision-maker with additional decisions and so on." Moreover, strategic consequentialists are sensitive to "the incentives that their choices create, to the informational limitations and asymmetries they face, and to the opportunities for misrepresentation these create, and also to considerations having to do with motives and intentions, since these are highly relevant to predicting how others will behave." (Woodward and Allman 2007, 185) In contrast, parametric

[96] The same distinction might be made for deontology, although Woodward and Allman do not explore this issue. The hardcore position by Kant on lying may count as an instance of parametric deontology. Parametric deontology includes absolute rights (and corresponding duties) that are not *prima facie* and must be followed without exception (*pereat mundus, fiat iustitia*). A deontological position according to which rights and duties are generally *prima facie* may count as a strategic form of deontological ethical system, since it allows the balancing of different moral claims and exceptions. I thank Bernard Baertschi for pointing this out to me.

consequentialism, whose main representative is Peter Singer, regards the behavior of other agents as independent of the agent's behavior, so that it stipulates away the considerations that strategic consequentialism takes into account[97]. So, contrary to Gigerenzer who proposes a wholesale rejection of maximizing strategies, Woodward and Allman do not go for a wholesale rejection of consequentialism, but attack parametric consequentialism only. In contrast, they are sympathetic towards strategic consequentialism and they link moral intuitions to it. Woodward's and Allman's neuromoral theory claims, among other things, that an improved knowledge of the machinery for moral judgments leads us to reject parametric consequentialism, but not strategic consequentialism[98]. This being said, the most interesting feature of Woodward's and Allman's neuromoral theory is that they explicitly address the meta-normativity problem. In the experiment by Dijksterhuis and colleagues (2006), the benchmark for the goodness of a choice was not a problem, since the cars were described with traits that were uncontroversially positive or negative (for a car). But how to find a similar normative standard for moral decision making strategies? Woodward's and Allman's attempt consists in defining, albeit in a provisional way, morally relevant information:

> [O]ne is more likely to arrive at a morally defensible decision/assessment of [action] A if this assessment reflects the operation of some process that exhibits the right sort of sensitivity and responsiveness to facts having to do with how oneself and others will be affected by A, how others are likely to respond to A, how this will affect all those concerned and so on – call this the morally relevant information. (Woodward and Allman 2007, 193)

Thus, the more morally relevant information a moral decision takes into account, the better. Still, this does not tell us whether the belief "One ought to rely on moral intuitions

[97] If this distinction is valid, then Greene's descriptive idea of consequentialism as a unitary psychological natural kind is likely to be wrong, since these two variants of consequentialism seems to rely on quite different cognitive processes, as strategic consequentialism must process a significantly bigger amount of data.

[98] It is unclear whether the distinction between strategic and parametric consequentialism maps in any significant way with the distinction between rule and act consequentialism.

and not to trust parametric consequentialism in morally relevant situations in which social learning is possible" is a moral judgment, a prudential judgment, or a judgment of some other sort. The nature of this 'ought' remains unexplained, even though at least we have been given a criterion that specifies under which conditions a moral judgment is better or worse than another. On the basis of this criterion, Woodward and Allman defend moral intuitions by claiming that, in many cases, they track morally relevant information. However, the parametric consequentialist could object that, at least in some cases, the information Woodward and Allman see as morally relevant is not morally relevant at all. To take for granted that some information is morally relevant seems to commit a *petitio principii* against the parametric consequentialist, as much as Sunstein was begging the question against a hardcore Kantian. If one says that the features mentioned by Woodward and Allman are important and that correct methods of moral decision making are those that track those features of the situation, then one is "stipulating away" parametric consequentialism, which regards those features as irrelevant in Woodward and Allman's reading. Similarly, parametric consequentialists "stipulate away" those features themselves from moral discourse and would say that correct methods of moral decision making are those that *do not* track those features of the situation (i.e. separateness of persons, integrity, etc). The two contenders seem to be on a par and the proposed criterion fails to establish a common ground between different moral decision making procedures. The core concept here is "morally relevant factor." I will deal with this concept below, in Ch. 4. Although Woodward and Allman realize that the problem of the criterion for the normative claims they make about moral decision making strategies is real and serious, the theory they sketch does not seem to possess the resources that are necessary to address it.

Finally, there is another positive consequence of an increased understanding of the machinery for moral judgments, a consequence that is not part of Woodward's and Allman's neuromoral theory because it concerns the meta-ethical level. Through increased knowledge, we can jettison misconceptions about what moral intuitions actually are.

Drawing on neuroimaging experiments, Woodward and Allman attack the idea that moral intuitions are linked to a domain-specific module or that they have to do with logical processing or visual perception. Hence, in these authors' opinion, philosophical views claiming that moral intuitions are connected with *a priori*, logical truths or with the perception of moral properties in the outer world are unlikely to be correct. But I mention this just to set it aside, as I cannot enter the topic of meta-ethical consequences of empirical investigation in this dissertation.

3.6. Walter Sinnott-Armstrong: the disunity of morality and the master argument

Philosopher Walter Sinnott-Armstrong puts forward two theories in moral psychology / philosophy. The former is the Disunity of Morality Theory. The latter is the Master Argument about Moral Intuitions. I examine them in turn.

The Disunity of Morality Theory is the view according to which a better understanding of the machinery for moral judgments makes us understand that the machinery for moral judgments itself does not exist as an object worthy of scientific investigation.

Sinnott-Armstrong's argument here runs like this:

Premise 1: what we normally label as 'morality' is a set of judgments that are not unified by any feature that enables universal generalizations.

Premise 2: what is bereft of features enabling universal generalizations is not a proper object of scientific investigation, as much as jade is not a proper object of mineralogical investigation because it is a compound of two minerals, jadeite and nephrite (to which mineralogical investigation applies instead).

Sinnott-Armstrong takes Premise 2 as uncontroversial and I believe he is right in doing so.

From these two premises stems the

Conclusion: we ought not to investigate "moral" phenomena, but to look at finer categories such as harm, fairness, sexual disgust, etc. that may have unifying properties of the kind required.

In other words, researchers should isolate smaller classes of judgments within one of these regions of "morality", present relevant scenarios from a single perspective, and look for the neural basis / evolutionary origin / psychological processes that correlate with that judgment.

This does not amount to a neuromoral theory, since Sinnott-Armstrong does not claim that expanded knowledge of the machinery for moral judgments would lead us to make better moral judgments. The normative consequences belong in this case to the domain of scientific methodology, i.e. have to do with how experimental moral psychologists should work.

Since Premise 2 is quite uncontroversial, everything hinges on Premise 1. Premise 1 amounts to the idea that moral judgments are a non-unified set that includes all moral judgments that intend to express or apply moral standards in contrast with legal, economic, aesthetic, prudential, etc. standards. The judger's intention is what makes the moral judgment moral, but in Sinnott-Armstrong's view the judgments that are coupled with this intention do not have any kind of unifying features that allow for scientifically valid generalizations. Sinnott-Armstrong's claim is in this case a negative existential: "There is no such a thing as a unifying feature of moral judgments that enables universal generalizations." (cf. Sinnott-Armstrong and Wheatley 2012). It is of course very difficult to prove a negative existential, but not all good arguments need to be definitive proofs, as Sinnott-Armstrong correctly believes. Sinnott-Armstrong tries to debunk the most notable possible candidates, saying that this suggests that no actual plausible candidate to the role of unifying property can (or could in the future) be located.

First, moral judgments are not based on harm, since a judgment like "It is morally praiseworthy to give a celebratory gift to your colleague who got tenure"[99] is moral but has nothing to do with harm.

[99] This example is from Sinnott-Armstrong and Wheatley (2012).

Secondly, the MFT is commendable as it underlines the variety of judgments that are intended as moral, but it is incomplete and vague. In particular Sinnott-Armstrong re-elaborates the objection by Suhler and Churchland (2011) and claims that the foundations could be many more than four, five, or six.

Thirdly, Turiel's (1983) characterization of moral violations as serious, harm-based, and authority independent clashes with some empirical results (Kelly et al. 2007) that show that some harm violations are authority dependent. For instance, 44% of participants said that a teacher is morally allowed to spank a student if the principal in turn allows her to do so, but only 5% of participants said that a teacher is morally allowed to spank a student if the principal in turn does not allow her to do so[100]. This shows that moral violations are not authority independent, or at least not with the kind of robustness Turiel and his followers took for granted.

Fourthly, Hare (1981) claims that all moral judgments are universalizable and prescriptive. Sinnott-Armstrong objects that moral judgments concerning the past, such as "It was morally wrong for Brutus to kill Caesar", and negative moral judgments, such as "It is not morally wrong to buy a lottery ticket", do not entail any kind of prescription.

Sinnott-Armstrong takes also other cases into account, but this brief review is sufficient to give an idea of his argumentative strategy. Having concluded his debunking, Sinnott-Armstrong puts forward a view in moral epistemology[101] called 'epistemic variantism' (Sinnott-Armstrong and Wheatley 2012, 373), according to which moral beliefs are justified by different methods or to different degrees in different areas of morality. Sinnott-Armstrong's argument is persuasive but not necessarily compelling. It is difficult to address this issue from the conceptual point of view only. What is more interesting for the

[100] This is the text of the scenarios: "Screen 1: It is against the law for teachers to spank students. Ms. Williams is a third grade teacher, and she knows about the law prohibiting spanking. She has also received clear instructions from her Principal not to spank students. But when a boy in her class is very disruptive and repeatedly hits other children, she spanks him. Screen 2: Now suppose that it was not against the law for teachers to spank students, and that Ms. Williams' Principal had told her that she could spank students who misbehave if she wanted to." (Kelly et al. 2007, 124)

[101] Moral epistemology is the discipline that deals with the justifications of moral judgments. In turn, a belief is justified for A when A thinks that the believer ought to have that belief on the basis of some epistemic ground.

purpose of this work is that Sinnott-Armstrong avails himself of empirical data to buttress his views concerning the Disunity of Morality. Schaich Borg and colleagues (2008) differentiated between three kinds of disgust (non-sexual moral disgust, i.e. toward gruesome violent acts, incest-related disgust, and pathogen-related disgust, i.e. the one caused by sipping one sister's urine) and showed that they elicit pretty different BOLD patterns in the human brain. But the most important study is by far Parkinson et al. (2011). Parkinson and co-workers parsed the moral domain into three domains: harm, disgust, and dishonesty. Then they wrote moral dilemmas that were ambiguous, in the sense that a pilot sample of participants were not making consistent judgments about the envisaged action or omission. Participants had to rate the action/omission as either morally wrong or not-wrong, i.e. in a binary way. This fMRI experiment revealed that the judgments "wrong" and "not wrong" did not correlate with any distinctive hemodynamic pattern, but that BOLD response was widely dissimilar for the three domains of morality. Harm violations were associated with regions that have to do with understanding and imaging actions, such as the L-DLPFC, the dorsomedial prefrontal cortex (henceforth DMPFC), the ACC, the p-STS, and the IPL. Dishonest transgressions recruited areas that are associated with Theory of Mind, non-mental representations, and the ability to simultaneously hold online multiple perspectives, such as the bilateral DMPFC, the TPJ, the IPL, and the L-DLPFC. Disgusting moral transgressions recruited social-emotional areas such as the amygdalae, the DMPFC, the ACC, the right temporal pole, and the PCC. Yet there is a cluster of voxels in the DMPFC that is consistently activated in all moral judgments. This area is traditionally associated with self-referential processes and to thinking about others in ambiguous circumstances, so that the authors conjecture that this area is processing ambiguity in this specific case. This empirical result suggests that no brain regions network common to diverse moral judgments can be found, but it is far from being definitive. A more systematic investigation of this kind ought to be carried out using clustering algorithms for

fMRI, possibly in a hypothesis-independent way. I know that Joshua Greene's lab is at the moment carrying out such an experiment, but the results are not out yet[102].

Sinnott-Armstrong's philosophical arguments are persuasive but not definitive, and the experimental evidence is again deeply suggestive, but not conclusive. Premise 1 in Sinnott-Armstrong's argument for Disunity is mainly empirical. It will be extremely interesting to see what experimental moral psychology will tell us about this in the next years and whether this young science will survive *its own* results. If Sinnott-Armstrong's Disunity of Morality Theory were supported by further evidence, experimental moral psychology should be deeply changed and carried out in a much more careful way, with more precise stimuli selection and attention at one specific domain at the time. There would be an experimental psychology of dishonesty, an experimental psychology of physical harm, and so on. Furthermore, from that moment on "morality" should be considered as a folk psychological term with no right of citizenship in empirical science. Arguably, the whole empirical investigation on morality would greatly change, but common beliefs about morality would not be altered, just as Einstein's general relativity theory has not substantially changed the folk concepts of time and space. According to the Disunity of Morality Theory, empirical results coupled with philosophical argument could leave a dent on scientific methodology and importantly alter the way in which experimental moral psychology is conducted.

Let us pass now to the other theory by Sinnott-Armstrong, the Master Argument about moral intuitions.

According to the Master Argument, whose full text is below, studies of framing effects and other psychological phenomena give humans reasons not to believe their moral intuitions *à la* Haidt (*all* of them).

In Sinnott-Armstrong's view, inference is any reasoning process that starts from one or more beliefs and is supposed to provide a justification for another belief. Moreover, a

[102] This is the research project of PhD Candidate Alek Chakroff in Greene's lab at Harvard, actually.

moral intuition can be really justified in two ways only: inferentially or non-inferentially. An intuition, which is a kind of belief, is justified inferentially if and only if it is justified because the believer can infer it from some other belief. A belief is justified non-inferentially if and only if it is justified independently of whether the believer can infer it from any other belief. If a belief is justified inferentially, the beliefs that justify it must in turn be justified. Moral intuitions are interpreted by Sinnott-Armstrong in the usual Haidtian way. Haidt's SIM must not be confused with Epistemological Moral Intuitionism (henceforth EMI), that is a position in moral epistemology. SIM is a completely descriptive theory, whereas EMI is a theory in moral epistemology which has to do with the justification of moral beliefs. According to EMI, some moral intuitions are justified non-inferentially in the sense described above. There are several methods of non-inferential justification normative intuitionists put forward, among which reflection, reliability, and others, but I do not need to give much detail on this here: the fact that these strategies for justification are non-inferential is sufficient for Sinnott-Armstrong's purposes. Of course, the eventual inexistence of non-inferential justification creates a regress, since if judgment A is inferentially justified by judgment B, then we need to understand how judgment B is in turn justified. If non-inferential justification does not exist, we have to appeal to judgment C in order to justify B, and so on *ad infinitum*. Hence, the inexistence of non-inferential justification brings grist to the moral skeptic's mill, since the moral skeptic in moral epistemology claims that no moral judgments are justified[103]. Sinnott-Armstrong is an outspoken Pyrrhonian moral skeptic (cf. Sinnott-Armstrong 2006) and so he would be

[103] The moral skeptic normally claims that no moral judgment is justified without qualifications that have to do with the contrast class that is taken into account in the justification process. However, the contrast class ought to be specified. For contrastivism, see Sinnott-Armstrong (2006, ch. 5). More specifically, a moral skeptic about extreme justification (i.e. justification out of the extreme contrast class) claims that no moral judgment is justified out the extreme contrast class that includes extreme views such as moral nihilism. In contrast, he claims that moral justification is possible out of a modest contrast class that includes commonsensical moral views only and excludes moral nihilism. According to Sinnott-Armstrong, no moral claim can be justified when confronting the moral nihilist without begging the question against him. The way in which, according to Sinnott-Armstrong, justification out of the modest contrast class is possible is essentially coherentist – cf. Sinnott-Armstrong (2006, ch. 10). However, in Sinnott-Armstrong's view, we cannot justify the claim that the modest contrast class, as opposed to the extreme contrast class and other contrast classes, is the *relevant* one for moral justification. This meta-skepticism about the relevance of contrast classes leads Sinnott-Armstrong to claim that moral justification (without qualification) is impossible.

very happy to show that non-inferential justification is impossible and, hence, EMI false. In order to do so, he marshals empirical evidence. His argument comprises 12 points and works in this way:

1) For any subject S, particular belief B, and class of beliefs C, if S is justified in believing that B is in C and is also justified in believing that a large percentage of beliefs in C are false, but S is not justified in believing that B falls into any class of beliefs C* of which a smaller percentage is false, then S is justified in believing that B has a large probability of being false.

(2) Informed adults are justified in believing that their own moral intuitions are in the class of moral intuitions.

(3) Informed adults are justified in believing that a large percentage of moral intuitions are false. (from studies of framing effects)

(4) Therefore, if an informed adult is not justified in believing that a certain moral intuition falls into any class of beliefs of which a smaller percentage is false, then the adult is justified in believing that this particular moral intuition has a large probability of being false. (from 1–3)

(5) A moral believer cannot be epistemically justified in holding a particular moral belief when that believer is justified in believing that the moral belief has a large probability of being false. (from the standard above)

(6) Therefore, if an informed adult is not justified in believing that a certain moral intuition falls into any class of beliefs of which a smaller percentage is false, then the adult is not epistemically justified in holding that moral intuition. (from 4–5)

(7) If someone is justified in believing that a belief falls into a class of beliefs of which a smaller percentage is false, then that person is able to infer that belief from the premise that it falls into such a class. (by definition of "able to infer")

(8) Therefore, an informed adult is not epistemically justified in holding a moral intuition unless that adult is able to infer that belief from some premises. (from 6–7)

(9) If a believer is not epistemically justified in holding a belief unless the believer is able to infer it from some premises, then the believer is not justified non-inferentially in holding the belief. (by definition of "non-inferentially")

(10) Therefore, no informed adult is non-inferentially justified in holding any moral intuition. (from 8–9)

(11) Moral intuitionism claims that some informed adults are non-inferentially justified in holding some moral intuitions. (by definition)

(12) Therefore, moral intuitionism is false. (from 10–11) (Sinnott-Armstrong 2008c, 99-100).

The impossibility of non-inferential justification is derived from the fact that experimental psychology results show that most of humans' moral intuitions are unreliable and hence unjustified. In presence of this widespread unreliability of moral intuitions, the only way in which a moral intuition A1 can be justified is by showing that, in A1's case, the conditions that cause moral intuitions to be unreliable are not active. But this amounts to a potential inference from "A1 has been made under circumstances that do not compromise judgments' credibility and hence falls into any class of judgments C* of which a small percentage is false"[104] to "A1 is morally justified". Hence, there is no other justification than inferential justification for these judgments, i.e. moral intuitions. As a consequence, EMI must be jettisoned. However, all this crucially relies on premise 3, which is empirical. Without premise 3 the argument cannot work. This is the point where experimental moral psychology comes in. Sinnott-Armstrong avails himself of the results by Kahneman, Tversky, and many others to show that in most cases humans are under one of these conditions when making moral judgments:

(1) partiality, i.e. the moral judgment affects the self-interest of the judge;

(2) moral disagreement;

(3) strong emotional states that cloud judgmental capacity;

(4) various kinds of cognitive illusions such as context illusions[105], generalization heuristics[106], and standard heuristics such as availability and representativeness;

(5) causal influence by unreliable or disputable sources.

[104] Of course, this belief ought in turn to be inferentially justified, and this makes us enter the regress Sinnott-Armstrong cherishes.

[105] For instance, to borrow Sinnott-Armstrong's example, being near a giant redwood makes a man appear small and being near a bonsai makes a man appear big.

[106] Such as the one according to which ovals that are shaded on the top appear concave (as a cave) and ovals that are shaded at the bottom appear convex (as an egg).

If any of these conditions obtains when a moral judgment is made, then the judgment can be justified only if it is confirmed by showing that we have some reason to believe that, in this specific case, the judgment is not influenced by these conditions, i.e. the judgment can be justified only through inferential justification. Sinnott-Armstrong claims that these conditions obtain, together or in isolation, often *enough* to justify Premise 3 in the Master Argument. Yet this latter claim is highly contentious. What is the meaning of "large" in Premise 3? It is not at all clear whether *enough* beliefs in the class of moral intuitions are false to make Premise 3 true. This is an empirical claim and it is almost impossible to test. Sinnott-Armstrong correctly specifies that any exact cut-off would be arbitrary[107]. However, since the proposed boundary is very vague and numbers are not available (what is the cardinality of the set of all existing moral intuitions?), it seems that Sinnott-Armstrong's quest to justify Premise 3 faces deep problems. How can humans decide whether a belief such as "Framing effects are pervasive *enough* to justify a general claim of unreliability for moral intuitions in absence of inferential confirmation" is correct? I see no way to do so. Therefore, judgment on this belief ought to be suspended. This belief cannot thus play any role in any argument, including Sinnott-Armstrong's argument, which falls in absence of its key Premise 3.

Summing up, Sinnott-Armstrong does not put forth a neuromoral theory, since an improved knowledge of the machinery for moral judgments, in his opinion, allows us to understand that most of human moral intuitions are false. However, this does not amount to making better moral judgments, but it amounts to making less moral mistakes. This would be possible by suspending judgment on our unjustified moral intuitions. Since in Sinnott-Armstrong's view it is very rare for a moral intuition to be correct, we would avoid lots of mistakes by stopping believing in unwarranted moral intuitions. This could thus be seen as a neuro-normative theory that recommends, on the basis of experimental results, widespread *epoché* about moral intuitions. This theory does not face the meta-normativity

[107] "I do not need or want to commit myself to any exact cutoff, since precise numbers are unavailable for moral beliefs anyway." (Sinnott-Armstrong 2008c, 101)

problem, because it does not differentiate between good or bad moral judgments. Yet it assumes too much from the available empirical results and takes experiments to justify a general claim on moral intuitions that they are likely never to justify.

Summing up the whole chapter, I hope to have persuaded the reader of the truth of the following two points:

(1) There are many descriptive alternatives to Greene's dual-process model and their explanatory plausibility is rather good, so that they compete on a par with Greene's ideas;

(2) There are many neuromoral theories and they regularly fall into the meta-normativity problem.

In the next chapter I examine Joshua Greene's neuromoral theory, arguably the most sophisticated so far, and by far the most discussed.

Chapter 4: Greene's neuromoral theory

In what follows I will grant *arguendo* that Greene is correct from the descriptive point of view, even though this is up for grabs, as I have tried to show in Ch. 2. Simply put, Greene's claim in normative ethics is the following: System2 ought to be favored relative to System1 in all cases of conflict between the two systems. Given the link Greene establishes between, on the one hand, the two systems and, on the other hand, two normative ethical theories (i.e. deontology and consequentialism), this means that consequentialism must be preferred to deontology in all cases of conflict. As he puts it: "[F]or me at least, understanding the source of my moral intuitions shifts the balance, in this case as well as in other cases, in a more Singerian, consequentialist direction" (Greene 2008a, 76). The position can be framed as a neuromoral theory: if we learn more about the machinery for moral judgments, we can make better moral judgments because we understand that we ought not to follow deontological intuitions when taking controversial moral decisions[108]. Yet, many details of Greene's normative theory need to be explained and arguments must be spelled out in a more careful way.

The first use of Greene's experimental results to argue for a normative claim in substantive ethics is by Singer (2005). Greene then built on Singer's paper to deliver his famous essay (2008a). Arguably, some of the arguments in Singer (2005) and Greene (2008a) are unclear and unconvincing. A step forward in the debate was made with the publication of a paper by Selim Berker (2009). Berker is critical towards both Singer and Greene. Even though I do not endorse all of the criticisms by Berker, the paper clarifies some of the important issues. Greene (2010), which is a response to Berker (2009), provides a

[108] This kind of neuromoral theory is structurally similar to the one put forward by Gigerenzer, with the important difference that the German psychologist argues that we should be wary of maximizing consequentialism and not deontology. Greene argues that improved knowledge about the machinery for moral judgment shows that deontology is unreliable because System1 often reacts to irrelevant features of a situation. Gigerenzer argues that improved knowledge about the machinery for moral judgment shows that maximizing consequentialism is unreliable because maximization is problematic due to the reasons highlighted in § 3.4.

substantially refined version of Greene's normative arguments. It also defends his experimental claims, but, as I have said, I will assume *arguendo* in this chapter that Greene's descriptive model is correct. Hence, I will mostly focus in what follows on Greene's response to Berker's paper and to the forthcoming paper by Greene. As a result of Berker's critique, Greene (2010) seems to have abandoned some of the arguments that could be attributed to him in his (2008a). In the first five sections (4.1. through 4.5.) I examine these arguments, namely the 'emotion good, reason bad' argument, and the Evolutionary Debunking Argument (henceforth EDA). In 4.4. I examine the Cultural Debunking Argument (henceforth CDA), which only Singer avails himself of. Then, in the central part of the chapter, from § 4.6. up to § 4.13., I concentrate on the AMIF, which is the main weapon Greene wields to fight his (and Singer's) anti-deontological battle. In the last part of the chapter, from § 4.14. to the end, I deal with other issues concerning Greene's neuromoral theory, such as his attack against the DDE and a possible *ad hominem* argument against the consequentialist theorist.

Here is a little preview of the main forthcoming attractions. I will claim that the AMIF, as it stands, is problematic because Greene assumes that some factors of a given scenario are morally irrelevant, whereas it is not at all clear that they are. Following Hare (1981), I will try to show that to say "Factor F1 is morally relevant" is equivalent to endorsing some moral principle that involves F1, such as for instance "Actions that possess factor F1 must be judged more leniently / more harshly than actions that do not possess it, *ceteris paribus*." Hence, the choice whether to consider a factor as morally relevant is as morally controversial as endorsing a moral principle. There are many moral principles in philosophy. Some are not problematic, for instance the following: "If you can achieve W by doing X or Y, and X involves inflicting physical harm to somebody, while Y involves no harm to anyone, then choose Y." In contrast, some are problematic, for instance the DDE, the DDA, and so on. Some claims Greene makes about the moral relevance of factors seem to be as controversial as these latter principles. For example, the role of

distance in morality seems to be very problematic. Some other claims Greene makes about the factors that affect people's moral intuitions are less controversial, as the ones about personal force. However, these claims are problematic in other ways, as I will point out below. Greene moves the trolley problem from principles to factors via empirical science, but in the end he gains little in doing so. Finally, Greene's neuromoral theory also falls into the meta-normativity problem, just as the other neuromoral theories I have examined in Ch. 3.

4.1. Singer and the 'Emotion bad, reason good' argument

Singer has been a staunch opponent of moral intuitions *à la* Haidt for a long time. His view is that that "all the particular judgments we intuitively make are likely to derive from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economical circumstances that now lie in the distant past" (Singer 1974, 516). Singer claims that Greene's results, together with Haidt's, show that moral intuitions cannot be taken seriously and must be substituted by rational moral theorizing. The main principle of rational morality is the equal consideration of the interests of all sentient beings (cf. Singer 1981, 102-103). Rational morality tells us to carry out the action that maximizes the expected furthering of those interests. Singer thinks that the experimental results buttress his claim against moral intuitions because they show that moral intuitions rely on emotions and are amenable to an explanation that does not require those intuitions to be true. In particular, it is possible to see Haidtian moral intuitions as adaptations that increased the inclusive fitness of ancestral hominid groups in the EEA, an environment that for sure "now lies in the distant past." For instance, humans' moral intuitions about consensual incest have increased the inclusive fitness of human groups in the EEA by preventing the birth of offspring with genetic diseases. There seem to be at least two arguments in Singer (2005). The former is the one

that Berker (2009, 316) dubs as the "emotions bad, reason good argument." The argument runs like this:

Premise 1: Deontological intuitions are driven by emotions, whereas consequentialist intuitions involve abstract reasoning.

Premise 2: Moral intuitions that are driven by emotions do not have any genuine normative force.

Conclusion: Therefore, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force.

This argument reflects the general rationalist stance that Singer takes towards moral philosophy and this is confirmed by the frequent reference to Kohlberg's theory of moral development in Singer (1981). Instead of starting from moral intuitions and using some form of RE to find a balanced set of CMJs, moral principles, and eventually background theories about human beings and society, we should, in Singer's opinion, follow a different path: "It seems preferable to proceed as Sidgwick did: search for undeniable fundamental axioms; build up a moral theory from them" (Singer 1974, 517) and again separate "those moral judgments that we owe to our evolutionary and cultural history, from those that have a rational base" (Singer 2005, 351). However, the "emotions are bad" argument begs the question against the proponent of the view that emotions help making better moral judgments[109]. For instance, Nussbaum (2001) holds such views. Singer provides no argument against the position of Nussbaum (and others). He seems to take for granted that rational information processing is better than emotional information processing at making "good" moral judgments or, equivalently, at tracking moral facts or the moral truth. But reasoning, as a decision making procedure, does not necessarily lead to "more desirable" or "more correct" decisions than following emotions. An argument is needed to establish

[109] I am referring to the idea that 'emotions are good', in the sense that emotions can be reliable guides to the moral good and help us to make good moral judgments. Apart from Nussbaum (2001), a similar stance is taken by Damasio (1994). This position ought not to be confused with sentimentalism, which I take to be a meta-ethical view. Namely, sentimentalism is the meta-ethical theory according to which moral judgments express (or are constituted by) emotions.

this. Perhaps reasoning brings about "better" decisions than following emotions under certain circumstances and "worse" decisions than following emotions under other circumstances. A philosophical theory about what counts as "better" in this context, plus empirical science, might tell us whether this is the case. A philosophical theory of this sort is necessary here because of the meta-normativity problem. It is not clear what is the normative standard according to which some decisions are "good" and some others are "bad." I mention this just to set it aside. Be as it may, emotions and non-maximizing cognitive processes are often multi-faceted, complex, and subtle, so that it is not obvious that they are bad guides for moral decision making. So emotions could in principle lead to "good" decisions, at least sometimes. This cannot be excluded *a priori.*

Secondly, Singer tries to use the 'emotion bad, reason good' argument to show that RE is a bad way of doing ethics. According to Singer, RE tries to systematize emotionally driven moral intuitions. However, this refers to Haidtian moral intuitions. The moral intuitions that enter RE are not intuitions of this kind, as I have explained in Ch. 1. Especially in WRE, where an equilibrium between CMJs about cases, general moral principles, and background theories about human beings and social institutions is sought, the moral judgments that enter the equilibrium process are both filtered, as they must be passed by impartial and competent judges that know very well about the case (cf. Rawls 1951), and revisable, since they can be modified at the margins when they do not fit with entrenched moral principles and background theories.

So the 'emotion bad, reason good' argument does not seem to work for CMJs.

Singer and Greene could argue that no matter how much filtering and balancing one does, the result of RE will be bad if one starts with "bad" moral intuitions. In relation to this, one can notice that the various forms of RE (and especially WRE) take all possible cautions to avoid "bad" moral intuitions. Furthermore, RE and WRE do not start with moral intuitions about cases only, but with moral principles, and in the case of WRE, with background

theories too. Singer[110] could further claim that the idea of starting with emotion-driven intuitions – even in the presence of filtering techniques – is *in principle* wrong, i.e. that the idea itself of RE is misconceived. However, Singer lacks a credible epistemology concerning his rational consequentialist intuitions. It is not clear how such can be distinguished from other general moral principles. I examine Singer's view about these consequentialist moral intuitions more in detail below, while speaking of the EDA.

A second objection Singer and Greene could raise, even if to my knowledge they do not actually do so, is that CMJs are idealizations, i.e. they are cognitive acts that people never (or very rarely) actually carry out in real life because the conditions for a belief to qualify as a CMJ are too strict. Rawls and followers, such as Norman Daniels (1979, 1980a, 1980b), might be seen as disconnected from experimental moral psychology because they have ideas concerning the formation of moral judgments that do not match with experimental results. An experimental philosopher would say that they have never left the armchair and ignored what came out of the lab. Yet, in order to condemn the upholders of RE and WRE in this way, experimental moral psychologists ought to reach some overlapping consensus about the workings of the machinery for moral judgments and persuasively show that CMJs are psychologically very rare. In other words, experiments should show that people are very seldom in the mental states described by Rawls (1951), because, contrary to Rawls's *desiderata*, actual moral judges often lack information, are biased, and so on. In this case we would have empirical evidence that just a few CMJs are passed in the actual world, so that it would be very difficult to locate some CMJs when the theorist starts the RE and looks for some moral judgments to feed into it as raw material. Yet we face here the same problem I highlighted while discussing Sinnott-Armstrong's Master Argument in § 3.6. When are CMJs experimentally rare *enough* to justify the claim that they do not actually exist, i.e. are idealizations? It seems hard to deal with this problem

[110] It is not clear whether Greene follows Singer on this, as in the last pages of (forthcoming) he advocates a form of "double-wide reflective equilibrium" that takes into account moral intuitions about cases, moral principles, background theories, and empirical facts about moral psychology.

of threshold arbitrariness. This problem aside, results in experimental moral psychology do not at the moment support the 'idealization hypothesis'. Empirical science has not reached that level of detail, and by far. So this remains a potential objection, which will have to wait for further progress in experimental moral psychology.

Greene would not necessarily share the 'emotions bad, reason good' argument, as he draws a distinction (see § 2.1. above) between alarm-like emotions and currency-like emotions. As I have written above, it is not clear whether this distinction has empirical backing and in general the role of emotion in Greene's model is not completely clear. Furthermore, Greene does not claim that following deontology is always wrong, but that it is wrong in cases of conflict with consequentialism and more in general about unfamiliar cases. Even though Greene is generally wary of emotional influence, this argument is more properly Singer's (2005) than Greene's.

4.2. The Evolutionary Debunking Argument: clarifying remarks

A second argument Singer uses, together with Greene (2008a), is the EDA. An EDA, according to Kahane (2011), has the following structure:

Causal premise: We believe that the evaluative proposition P1 is true because we have an intuition that P1, and there is an evolutionary explanation of our intuition that P1;

Epistemic premise: Biological evolution is not a truth-tracking process with respect to evaluative truth.

Conclusion: Therefore, we should not believe that P1.

Singer limits the scope of the argument to deontological intuitions.

In this section I will make some remarks that aim at making the discussion clearer. These considerations are not objections to the EDA or to Singer's view. I will expound my argument against Singer's use of the EDA in the following section. To repeat, the EDA is the second argument that Singer and Greene use to attack deontology. I have already

examined the 'emotions bad, reason good' argument. I will examine the AMIF in the sections below.

First, many judgments in normative ethics, such as "it is morally wrong to kill an adult member of your in-group that has not tried to assault you," are amenable to an evolutionary explanation. Given that humans are hyper-social animals and that human groups competed for resources in the EEA, evaluative judgments that decrease intra-group violence and free-riding are likely to increase group fitness in inter-group competition[111]. These judgments can be explained via biological evolution. Other moral judgments require cultural mechanisms to be explained. For instance, "It is morally mandatory to help the disabled 70-year old woman to survive while the tribe is migrating to the south" *cannot* be explained through biological evolution only. As the elderly woman consumes resources, produces none, and cannot reproduce, the tribe is very likely to be better off without her, except under specific circumstances (e.g. the woman has some kind of technical expertise that she could transmit to other members of the group; the woman is endowed with some social role that makes her survival important for group cohesion and functioning, and so on). So it makes little sense, in the logic of biological evolution, to help the both sterile and weak – they just consume precious resources that could be spent in a more efficient way. But, importantly, this judgment *can* be explained through *cultural* mechanisms. For instance, one can imagine that, in a human group in the EEA, most members would be kin. Hence, human traits such as linguistic accent or ways of clothing would become proxies for kinship. The use of these proxies to distinguish kin from non-kin would lead to the formation of norms that promote helpful behavior toward those that exhibit the proxy trait, The above judgment concerning the elderly woman could be one of such judgments. I argue that the Axiom of Benevolence, deemed by Singer (2005) and by de Lazari-Radek and Singer (2012) an *a priori* truth, is likely to have evolved in this way, starting from

[111] As to the debate relative to group selection, see § 2.1.

solidarity norms that are a cultural extension of helping behavior toward kin, but more about this can be found below.

Secondly, causal explanations are not justifications. As Tersman (2008) pointed out, explaining why subject S1 holds the belief that P2 is true is different from explaining why it is true that P2. In some contexts, knowing how a belief was formed tells us little about whether the belief is epistemically justified or whether it is true. German chemist F. August Kekulé conjectured the structural formula of benzene starting from a day-dream in which a snake bit its tail, but the unreliability of the belief formation process does not indict the belief itself. The reverie was the cause of the belief, but it does not justify the belief, nor does it show that the belief is unwarranted. In this specific case, the justification depends on the epistemic standards of the empirical science known as chemistry and these standards are independent of how the belief was caused. If the belief about the structure of benzene helps explain experiments and data points that concern benzene, then this belief will enjoy an increasing degree of epistemic justification inside the empirical science it belongs to. Even though explanation and justification are normally uncoupled, there are some causal explanations that make justification of some beliefs unlikely. In contrast with empirical beliefs[112], evaluative beliefs can be justified in many different ways: there are coherentist justification methods, that start from evaluative principles and evaluative intuitions already present in a social group (for instance, RE and WRE), then various forms of evaluative intuitionism of which EMI is the moral variant, various forms of moral naturalism that link evaluative justification to a set of natural facts, Walter Sinnott-Armstrong's (2006, Ch. 5) contrastivism, contractarian views such as Scanlon's (1998), and many others. So different theories of justification of evaluative beliefs exist, but I mention this just to put it aside. Be as it may, many (but not all) positions in the epistemology of evaluative judgments assume that, in order for moral judgments to be

[112] I do not need now to enter the details concerning which beliefs are amenable to a debunking explanation and which are not. It is sufficient to note that evaluative beliefs seem to be beliefs that may fall prey to EDAs.

justified, there must be a tracking relationship between evaluative judgments and their factual grounds[113]. These grounds are constituted by alleged moral, aesthetic, political, economic, (add your evaluative domain of choice here) facts[114]. If it can be shown that evaluative belief P2 is caused by some causal process that is incompatible with the perception of such facts, then the evaluative claim seems to be debunked. Judgments that do not track the corresponding evaluative facts can be labeled as 'off-track' and they instantiate a 'tracking failure' (Lillehammer 2010, 365). The existence of evaluative facts is of course controversial. However, I will assume here that the concept of tracking, that requires the concept of evaluative facts, is important for the justification of evaluative judgments, among which moral judgments. If lack of tracking shows that justification is impossible, then some explanations of evaluative judgments, such as those that refer to biological evolution, can be seen as debunking.

Nozick (1981, 346 ff.) interestingly tries to show that such explanations, which he calls 'invisible hand explanations'[115], are *not* debunking. In order to do that, he uses two arguments.

The former is an analogy with mathematics[116]. Nozick correctly points out that humans know nothing certain about mathematical facts – there is no agreement whatsoever on what they are, exactly as there is no agreement on the existence and eventual properties of evaluative facts. However, this does not lead us to reject mathematical statements. Hence,

[113] "It is widely, if not universally, accepted that tracking failure would impugn the epistemic credentials of our ethical beliefs" (Lillehammer 2010, 365)

[114] The notion of 'evaluative fact' is complex. I do not want to explore it in depth here. I assume that there are evaluative judgments, such as political judgments (e.g. 'It is politically useful to do X'), economic judgments (e.g. 'It is economically wise to do X'), aesthetic judgments (e.g. 'It is beautiful to paint in such and such a way'), moral judgments (e.g. 'It is morally praiseworthy to give all of one's money to the poor'), and so on. Evaluative facts are the alleged facts that make these judgments true or false under a cognitivist and realist interpretation of these judgments. I do not include social, cultural, or evolutionary facts under the label of 'evaluative facts'.

[115] An invisible hand explanation shows "how some overall pattern, which one would have thought had to be produced by a successful attempt to realize the pattern, instead was produced and maintained by a process that in no way had the overall pattern 'in mind'." (Nozick 1974, 18)

[116] Albeit in a different way, this analogy is also suggested by de Lazari-Radek and Singer (2012). These authors think that both the capacity of identifying moral truths and the capacity of elaborating abstract mathematical theorems are not evolutionarily useful, but have evolved as a by-product of human reasoning capacity, that, contrary to the aforementioned capacities, could increase the inclusive fitness of the ancestors of man in the EEA and hence be selected for.

doubts about the corresponding facts are not sufficient to debunk the justification of some beliefs. If the analogy between mathematical judgments and evaluative judgments holds, then, in Nozick's opinion, doubts about evaluative facts are not enough to debunk the justification of evaluative beliefs.

To this argument I reply that there is a significant *disanalogy* between mathematical beliefs and evaluative beliefs. As Kitcher (2005) and Joyce (2006, 184) correctly note, there is a fundamental distinction between arithmetic and evaluative domain. Arithmetic capacity increased inclusive fitness in the EEA because the beliefs it produced tracked some empirical truths, such as "there are three gazelles, two on the left among the bushes and one on the right under the shadow of the big tree." Arithmetic capacity seems to be an evolved ability that tracks empirical truths. Its being caused by biological evolution does not exclude tracking. In a similar way, the evolved abilities of *H sapiens* specimens to perceive and handle mid-sized objects were fitness-enhancing, have evolved because they were tracking some empirical truths, and their being caused by evolution does not exclude their ability to track relevant facts. In contrast, evaluative judgments increased inclusive fitness because they were believed to be true, but their actual truth was not required to make them fitness-enhancing. Thus, evolutionary explanations of empirical judgments include their truth, whereas evolutionary explanations of evaluative judgments do not include their truth. Hence, evolutionary explanations of the former judgments are not debunking, whereas evolutionary explanations of the latter could (at least) be.

According to the latter argument by Nozick, evaluative judgments have been enriched, along human evolution, with new cognitive abilities such as language, writing, and increasingly abstract thinking. Hominids have become rational. Hence, even assuming *arguendo* that moral judgments were *not* tracking moral facts *in the EEA*, it is well possible that they are tracking moral facts *now*, thanks to the added abilities. Even if primitive

forms of morality, such as kin altruism[117] and reciprocity[118], might have evolved simply due to the positive effects on fitness of the behaviors they favor, succeeding forms of morality might have acquired a link to human reasoning that would transform them into something substantially different from what morality used to be in hominid groups[119]. These new forms of morality could track moral facts even though the preceding forms of morality were not.

I am skeptical about Nozick's argument. If there is an 'invisible hand explanation' for a given evaluation, it would be a miracle for the evaluation to track an evaluative fact[120], since the evaluation is causally driven by factors that have nothing to do with such facts. If evaluative judgments were tracking selective pressures in the EEA and not alleged evaluative facts, it is not clear how the appearance of higher capacities would necessarily modify this causal link and transform off-track evaluative intuitions into on-track evaluative intuitions. We need empirical data to understand what the import of these new abilities (i.e. rational abilities) on tracking is. Besides, these data will not be easy to gather, since these abilities have appeared in a distant past.

Concluding, I think that the proponent of the EDA is correct in asserting that some kind of explanations, such as invisible hand explanations, are a problem for the justification of evaluative claims. As this kind of explanation makes the tracking of the relevant facts rather implausible, and it is commonly assumed that tracking is a necessary condition for justification, evaluative judgments that are amenable to explanations of this sort seem to be difficult to justify. EDAs do not prove that evaluative facts do not exist, yet evolutionary explanations make their existence quite unlikely. In my opinion, since they are not required for explanations, these alleged facts would quickly fall prey to the celebrated blade of friar William of Ockham. It is not clear why there should be evaluative facts if evaluations can

[117] For this form of altruism, see Dawkins (2006/1976, 94)
[118] For this form of altruism, see Trivers (1971).
[119] This argument could be useful to defend the position of Singer, that I attack below in § 4.3. However, to my knowledge Singer does not use this argument.
[120] For a similar view, cf. Street (2006).

be explained by processes that have nothing to do with evaluative facts. The amount of offspring that survive until reproductive age in a given hominid group could be one of such processes. Hence, an argument that starts with evolutionary causality and ends with anti-realism about evaluative facts, such as the one crafted by Street (2006), is at least plausible.

4.3. The Evolutionary Debunking Argument: the case against Singer

Singer aims at concluding that:

(1) a general EDA along the lines of Street (2006) and Joyce (2006) does not hold,

(2) it is possible to target the EDA against deontological moral judgments.

On the contrary, I see no ways to use the argument in a limited way. As Levy (2007), Tersman (2008), Kahane (2011) and others have already pointed out, the consequentialist tenets cherished by Singer and Greene seem to be as amenable to an EDA as deontological tenets. I also think that consequentialists could have a lot to lose from EDAs.

Two parts of a typical consequentialist theory can be attacked.

First and foremost, consequentialism requires an account of well-being in order to decide which consequences are good and which are bad. Utilitarian variants of consequentialism require this account even more than non-utilitarian forms of consequentialism because it seems to be very complicated to make sense of utility in absence of elucidations on well-being. Claims about well-being, such as "It is good to avoid pain" or "It is good not to get sick", are evaluative claims[121]. However, they are easily explained through biological evolution. Organisms that do not believe that pain ought to be avoided make fewer efforts to avoid it than organisms that believe that pain ought to be avoided. Hence, organisms of the former kind usually compromise their physical integrity and die early[122]. Organisms that are diseased face lower chances of reproduction, decreased coping with environmental

[121] I assume this. I do not think that the claim is actually controversial.

[122] Of course, not all organisms avoid pain *because* they believe that pain ought to be avoided. Indeed, most metazoans do not reach any plausible threshold for belief attribution. Most metazoans, such as for example a specimen of *Drosophila melanogaster*, avoid pain because of blind neural mechanisms. Nonetheless, avoiding pain still produces positive effects for inclusive reproductive fitness.

challenges, and increased probability to die. There is no need of specific prudential facts in order to explain these claims. Empirical facts are sufficient to provide an explanation. So, if one admits that an EDA undermines justification for other kinds of evaluative claims, it is difficult to say why it should not work for these evaluative judgments about well-being. In other words, if the EDA works in general, then it debunks claims about well-being too. In order to reply to this objection, de Lazari-Radek and Singer (2012, 28) write that "we will limit ourselves to pointing out that if no theory of well-being or intrinsic value were immune to a debunking explanation, this would show only that no theory could be preferred over others on the ground that it alone cannot be debunked. It could not show that no theory of well-being is true." In other words, what de Lazari-Radek and Singer are claiming is that, even if an EDA against judgments about well-being works, it does not show that they are false, but simply that they cannot be put into a ranking order relative to their resistance to EDAs, as all of them have been debunked. This is true, but it is not sufficient for the consequentialist theorist. If a successful EDA for well-being judgments were available, these judgments ought to be suspended and declared neither true nor false. Pyrrhonian skepticism about well-being would ensue and no forms of consequentialism (let alone utilitarianism) would work. If the justification of a claim is debunked, then we have no reason for believing that it is true and no reason for believing that it is false. In this situation we ought to suspend judgment. And no version of consequentialism can work if claims about well-being are under *epoché*.

Secondly, also consequentialist "ethical axioms" (Singer 2005, 351) or Sidgwickian moral intuitions[123], that count as the fundamental tenets of consequentialism, are potentially vulnerable to EDAs. A group of people that endorse "It is *ceteris paribus* morally better to save more human lives from death than less" will be advantaged relative to a group that does not partake this principle, since the latter group will incur in avoidable losses of human lives that could have been prevented and that imperil the standing of the group in its

[123] See Ch. 1, above.

competition with other human groups for the access to resources. Hence, if the EDA works in general, then it works for (at least some) rational moral axioms too[124].

Singer's response to this criticism (de Lazari-Radek and Singer 2012) hinges on the Axiom of Benevolence. In agreement with Singer, I think that the endorsement of this principle is widespread in many cultures and the principle is often taken to be a fundamental moral claim. But contrary to Singer, I think there is a more plausible explanation for this than the one he provides. De Lazari-Radek and Singer (2012) argue that an EDA for the Axiom of Benevolence is not available. They namely think that this axiom is a rational, *a priori* truth. It is true that believing in the Axiom of Benevolence does not seem to increase the inclusive fitness of the individual believer. Even accepting some form of group selection, evolutionary pressures seem to recommend nastiness toward the out-groups and non-human animals. This point gives credibility to Singer's idea that no EDA for the Axiom of Benevolence is possible. However, in order to resist this idea, it could be argued, following Kahane (2011, 119), that universal altruism is a reasoned extension of altruism towards kin and in-group, that an EDA is available for the latter form of altruism, and that, as a consequence, the EDA also strikes the universal form that has come out of the limited form. In other words, there are forms of altruism that were favored, under certain conditions, by biological evolution in the EEA, for instance altruism towards kin and reciprocal cooperation[125]. However, broader forms of altruism are created by generalization of these primitive and more limited forms through the use of human rational faculty. Hence, these more ancient forms are essential for the more general ones to develop. But, as Kahane points out, there is an EDA against those primeval forms. Since these are the necessary bases of the more modern forms, their debunking also leads to loss

[124] Another principle that can be considered is the following: "It is *ceteris paribus* morally better to save more human lives from death than less within the in-group." If we stick to Sidgwick's (1907) idea of philosophical intuition, this is not a rational moral intuition, since it is in contrast with the Axiom of Benevolence. However, it may have evolved in the EEA too and it is vulnerable to an EDA as much as its rational counterpart. I think that the transition from the principle quoted here to its rational, impartial version has been caused by cultural forces, as I explain below.

[125] In reciprocal cooperation organism A helps B so that B will in turn help A.

of justification for the newest ones. I add that a notable Christian principle, i.e. the idea that all human beings count as brothers and sisters, embodies pretty well from the cultural point of view the extension of kin altruism into the idea of universal benevolence[126]. Hence, there is cultural evidence of the extension of altruistic behavior Kahane's argument hinges on.

Singer replies to Kahane's challenge by claiming that we believe in the Axiom of Benevolence because it stems *a priori* from human beings' rational capacity and the human capacity of finding such an *a priori* moral truth is a by-product of a more general human rational capacity. So Singer does not connect the Axiom of Benevolence and the impartial, universal altruism it implies to preceding forms of altruism. This fits well with his overall rejection of Haidtian moral intuitions. Hence, the Axiom of Benevolence is uncoupled, in Singer's view, from forms of altruism that humans share with primates and that are amenable to eventual EDA. According to Singer, it derives from the structure of rational justification itself: "the element of disinterestedness inherent in the idea of justifying one's conduct to society as a whole, and extending this into the principle that to be ethical, a decision must give equal weight to the interests of all affected by it" (Singer 1981, 100).

Of course, there are moral systems that do not necessarily include this Axiom. There were cultures, such as the ones embodied by the Nazi party in the 1930s and early 1940s[127] and by Stalinist communism, that considered a moral duty, and sometimes a burdensome moral duty, to kill the members of other human groups, such as Jews, Gypsies, communists, homosexuals, class enemies, deviationists, and so on. There were cultures in which slavery was ordinary and expending the lives of slaves for fun was no problem at all[128]. A memorable article by Bennett (1974) describes the sense of guilt that Huckleberry Finn, the well-known fictional character by American writer Mark Twain, experiences when he frees

[126] I owe this point to Bernard Baertschi.
[127] On Nazi morality, cf. Glover (1999, ch. 37)
[128] For moral problems relative to Roman arenas, see Hare (1981, 142).

an African American slave from his owner. Huck thinks he has wronged the owner and feels very badly about his act.

Singer is well aware of this. He thinks that the Axiom of Benevolence, in his opinion embodied by well-known principles in important ethical traditions such as Christianity, Confucianism, Hinduism, and Buddhism (de Lazari-Radek and Singer 2012, 26), is the necessary and *a priori* end of a process of moral progress whose steps are dictated by historic contingencies, but whose direction is fixed as long as human rational capacity is not fundamentally altered. As long as human beings have the brains they now have, Singer thinks, there will be a slow moral progress that will ultimately lead to the equal consideration of the interests of all sentient beings. Admittedly, the possibilities of falsifying such a hypothesis are tiny. What kind of empirical evidence can be marshaled to disconfirm it? At the same time, Singer provides little positive data to buttress this hypothesis.

I provide here an alternative and more plausible explanation of the belief in the Axiom of Benevolence and the belief that it is a fundamental ethical tenet. It can be hypothesized that this "axiom" has appeared due to contingent cultural causes, and not because it tracks or represents some rational and *a priori* moral fact. So I provide here a cultural explanation for the Axiom of Benevolence. The need for a morality that facilitated the cooperation of unrelated individuals in urban societies in which trading and market relationships were increasingly common and the boundaries between in-group and out-group were becoming more and more blurry could provide a reasonable account of the appearance of the Axiom of Benevolence, assuming *arguendo* that it actually enjoys a high degree of cross-cultural consensus[129]. If this is a reasonable explanation, as I think it is, the Axiom of Benevolence could have appeared to enhance human cooperation in urban societies. Similarly, preceding systems of norms, such as the very widespread norms against physical harm, were already bringing about the same effect in small-scale, hunter-gatherer society. Hence,

[129] For a similar cultural explanation of the Axiom of Benevolence, see Tersman (2008).

the Axiom would roughly play the same function as those earlier norms. The Axiom would result from a slow expansion in the number of those that were entitled for moral protection and status. Reciprocal altruism got extended into harm norms valid for the in-group. More inclusive norms came in as society became more complex, anonymous, market-based, and large. This is of course just a hypothesis, but it has the advantage of parsimony relative to Singer's, which invokes rational moral facts[130]. If this hypothesis is correct, it is possible to run Kahane's argument: if universal benevolence is a cultural extension of norms that have originated from limited altruism and reciprocity, and the EDA hits moral beliefs linked both to reciprocity and to limited altruism, then universal benevolence may be in trouble. It must be underscored that the EDA *cannot* show that the Axiom is false. It can show that it is unjustified, which is different. In other words, if the EDA can indirectly reach humans' belief in the Axiom of Benevolence, what we ought to do it is to put it under *epoché*. The Axiom might still be true, but we have no reasons to think so. It is sometimes the case that we hold true beliefs even though we have followed an unreliable and wrong method to acquire them. For example, I can learn at 8 pm that it is 8 pm from a non-working clock that *just happens* to be displaying 8 pm. So I do not claim that, if Singer does not succeed in shielding the Axiom from an EDA, then the Axiom is false. Pyrrhonian skepticism about the Axiom of Benevolence would follow.

Besides, even assuming *arguendo* that this Axiom *is* immune, the utilitarian theorist would still have big problems to defend her standard claims if EDAs were valid, since EDAs undermine other important tenets of this normative ethical theory and, especially, accounts of well-being. For example, it undermines the claim that it is better to save more human lives rather than fewer.

What precedes concerned the EDA as presented by Singer. I have shown that if the EDA works, it can potentially cut down all normative ethical theories and to lead to moral anti-realism (as argued by Joyce 2006 and Street 2006). Singer wants to use the EDA in a

[130] Starting from Mackie (1977), a lot could be said about this issues, of course. But I do not want to enter deep metaethical issues here, as they could lead the thesis away from its focus on Greene.

targeted way, but it is not clear how he can stop the argument from affecting the moral principles he holds dear and sees as fundamental. If the EDA works (and I do not take a stance on this), then it hits both globalist and very general moral intuitions, such as the ones of the act utilitarian, and local moral intuitions about cases that are important for deontological theorizing.

4.4. The Cultural Debunking Argument

Before passing to Greene's recent positions, I would like to consider another kind of debunking strategy that Singer employs, the CDA. I claim that CDAs are different from EDAs and that they are worse. The CDA runs like this:

Causal premise: We believe that the evaluative proposition P1 is true because we have an intuition that P1, and there is a cultural explanation[131] of our intuition that P1;

Epistemic premise: Cultural history is not a truth-tracking process with respect to evaluative truth.

Conclusion: Therefore, we should not believe that P1.

To my knowledge, Singer does not spell out the argument in an explicit way in any of his texts. So this is my rendition of the argument. If it is difficult to assess the soundness and scope of the EDA, the CDA raises fewer problems, as it is a weak claim. Singer seems to consider this argument as valid. His quote above about "warped views on sex and bodily functions" seems to point in this direction. The epistemic premise of this argument seems to be highly questionable. In opposition to biological evolution, that has in principle nothing to do with moral facts or how humans ought to live, cultural history could track the moral truth. If Singer himself is right about his rationalism and his idea of moral progress, cultural history looks like a process of slow but constant nearing to the moral ideal embodied by the Axiom of Benevolence. Furthermore, there are many normative ethical

[131] For the present purposes, a cultural explanation obtains when a moral intuition can be explained by reference to some religious tenet or moral belief that has been culturally transmitted in a given human group across history.

systems that *tried* to spot the moral truth. Hence, either there is a good argument leading us to believe that they all fail, or a *prima facie* claim that cultural history cannot track moral facts is unwarranted. Perhaps some cultural traditions have got it right and have grasped some moral facts – it might be Tibetan Mahayana Buddhism, or Mormonism, or act Utilitarianism, or you name it. If a given moral claim is explained by a certain cultural tradition, it might be the case that the tradition is right and has pin-pointed a moral truth, since (at least some) cultures make conscious efforts to locate moral truths. Of course it may also be the case that a certain cultural tradition has failed to track moral facts even though its members think that they have grasped them and see themselves as holding the moral truth. But this need not concern us here: the point is whether a cultural tradition *can* grasp the moral truth, not under which conditions it actually *does so*. I am not interested here in discriminating between successful and unsuccessful moral cultures. I am trying to understand whether cultural explanations are debunking. It seems that, irrespective of whether such cultural beliefs (that were created by some human group in some contingent historical circumstances) *successfully* track the moral truth or not, a reference to the efforts of cultures to form moral beliefs in an explanation of a moral claim does not undermine the justification of such claim. There will be some cultural processes that grasp some kind of evaluative truth and preserve it through time, and some other cultural processes that do not do so.

We can apply here the point by Fine (2006) and Levy (2006b) about Haidt I mentioned in § 3.1. It may be the case that an individual, I1, learns an evaluative belief from someone else, I2. Suppose that I2 has in turn acquired the evaluative belief in a non-cultural way that guarantees tracking relative to the relevant evaluative facts, e.g. personal experience or reasoning or you name it. In this case, the facts play no role in the explanation as to *why* I1 holds that belief. However, the evaluative belief might still be true and justified because of I2's acquisition. It is likely that I1 cannot justify her belief, but this does not rule out the

hypothesis that the evaluative belief is true and justified. This is another reason to maintain that cultural learning does not exclude truth tracking relative to alleged evaluative facts[132].

Another difference between biological and cultural mechanisms is that biological evolution makes no reference to moral facts, whereas cultural traditions in moral thinking do, at least sometimes. I see no reason why cultures should fail to track moral facts, assuming *arguendo* that these exist. Hence, EDAs might work (again, I am not taking a stance on this), but CDAs lead nowhere. In my opinion, a cultural explanation is simply irrelevant for the eventual truth of a given moral belief. If "It is morally forbidden to eat lobster" is explained by a reference to Hebraism, then this does not show that this moral belief is right, as much as it does not show that this belief is wrong. So I argue that the whole idea of CDAs is misconceived. Therefore, Singer has a weapon less to strike the deontologist.

On the same topic, it could be argued that Singer's position on the CDA is self-defeating. Tersman (2008, 402) correctly notices that the Axiom of Benevolence Singer cherishes so much might be explained by reference to a cultural tradition. A part of the axiom states that "The good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other." Tersman pointedly remarks that "replace «Universe» with «God» and you get a doctrine that will impress many a Christian." Sidgwick's views are easily explained by England's Christian heritage. If the Axiom of Benevolence exists outside Christianity, this explanation is insufficient, but in the specific case of Sidgwick this explanation might be correct. Only historians of philosophy could tell us. Therefore, if CDAs are actually debunking, the Axiom of Benevolence (at least insofar as Sidgwick's version is concerned) is doomed. Singer wants to use CDAs to attack the deontologist and get rid of traditional views in ethics. However, the argument cuts both ways, since it explains the bases of Singer's moral system away. So, even assuming *arguendo* that Singer's CDA works, it works too much for Singer's purposes.

[132] I owe this point to Matteo Mameli.

4.5. Greene on the Evolutionary Debunking Argument

Greene (2008a) does not use the CDA[133] and, as far as the EDA is concerned, he runs on roughly the same lines. Nonetheless, Greene provides two arguments that make his position slightly different from Singer's. I will now discuss these arguments.

The former argument (2010, 19) distinguishes philosophical intuitions, that he describes as "natural, untutored judgments," and psychological intuitions, whose driving mechanisms are not accessible to consciousness and that depend on System1. Of course, Greene claims that deontological intuitions such as the one against consensual incest are psychological and that consequentialist intuitions such as the Axiom of Benevolence are philosophical only.

In the latter argument, Greene (forthcoming, § VI) says that consequentialist intuitions are 'philosophical' (in Sidgwickian terms[134]) and not 'perceptual' or 'dogmatic', i.e. they are very general. They are not linked to particular cases, so that they seem, in Greene's opinion, to have to do with neither evolved evaluative responses to specific kinds of situations nor raw gut feelings that are tightly connected with emotions.

The long and the short of both arguments is that consequentialist intuitions have not been caused by evolutionary dynamics in a direct way, so that they cannot be targeted through an EDA. The distinction that Greene draws is similar to the point by Levy (2006a) that consequentialists are 'globalists' and deontologists are 'localists'. Consequentialist theorists start from very general claims about the nature of the good and derive from those, through detailed empirical descriptions, moral judgments about cases. In contrast, the deontologist looks for principles that account for already formed moral judgments about

[133] Greene is silent on this, so I do not take him to share Singer's view on this point.

[134] It is interesting to notice that Sidgwick spends some words on debunking explanations of moral claims, that he dubs as attacks on "psychogonical grounds" (1907, 213). Sidgwick claims that the argument cannot indict the fundamental concepts of ethics, but that moral intuitions must be checked in order to control whether they have been the result of a biased psychological process.

cases[135]. As Levy (2006a) correctly noticed, a global moral intuition is no less a moral intuition than a local moral intuition. As I have written above, it is not obvious and it is at least disputable that consequentialist moral intuitions are immune to EDAs, as Greene seems to assume. In particular, the Sidgwickian idea that consequentialist moral intuitions are more general than those used by other normative ethical systems does not seem to confer advantages to the consequentialist theorist relative to the EDA. Being general or abstract does not have any link with the EDA as an argument, nor is it a factor that makes the EDA less cogent, if it is cogent at all. In order to assess his claims on this topic more carefully, Greene ought to give more details about his ideas on consequentialist basic tenets. I hope he will expand this important topic in his upcoming book.

Summing up, Singer and Greene do not seem to be justified in using the EDA in a targeted way against deontology, since they cannot differentiate epistemologically between the rational intuitions they want to defend and the emotive intuitions they want to abandon[136]. However, I have not provided here an argument that shows that it is *in general* impossible to restrict the EDA to some moral intuitions without affecting all moral intuitions. This would require delving into complex meta-ethical issues and this would lead this thesis out of topic. My claim is solely that the kind of restriction of the EDA Singer and Greene advocate seems to be deeply problematic because they fail to provide epistemological details about their cherished 'consequentialist intuitions'. It is thus very unlikely that Singer and Greene can *at the moment* make use of this piece of philosophical machinery to further their sheer consequentialist agenda. I do not want to take a stance here about both the general validity of EDAs and the validity of EDAs in the moral domain more

[135] There are actually two different distinctions: Greene distinguishes general intuitions (that are not selected for by biological evolution) from particular intuitions (that are selected for by biological evolution). Levy distinguishes 'globalist' and 'localist' moral intuitions. However, these two distinctions seem to map on each other in Greene's opinion: globalist intuitions are general and localist intuitions are particular and about cases.

[136] Singer and Greene might succeed in doing so in the future, but they fail to do so at the moment. There is no reason, in principle, to deem this endeavor hopeless, though.

specifically. Again, this would lead me out of the topic of this text and into deep meta-ethical problems.

So both the EDA and the 'emotion bad, reason good' argument do not give much traction to Singer's and Greene's neuromoral claim. To repeat, Singer's and Greene's neuromoral claim is that a better understanding of the machinery for moral judgments leads us to be wary of Haidtian moral intuitions and hence of the normative theory (deontology) that is allegedly based upon them. The most interesting argument to buttress this neuromoral claim is the AMIF.

4.6. The Argument from Morally Irrelevant Factors (AMIF) on personal force: introduction

Greene (2010, forthcoming) uses one general argumentative structure to derive normative conclusion from empirical facts and normative assumptions. The argument may be made clearer through an example made by Greene himself (forthcoming, § IV), even though this example is not specifically linked to Greene's experimental work.

Empirical premise: decisions of capital punishment are sometimes affected by unconscious racist biases.

Normative premise: capital juries ought not to be affected by unconscious racist biases.

Normative conclusion: capital juries sometimes make wrong decisions.

The conclusion implies that these juries sometimes take decisions that are different from those they ought to take. It follows that such juries ought to try, or to try harder, to avoid being influenced by racist biases.

From this structure Greene draws the AMIF, that in its last version (2010) functions in the following way:

Empirical premise: moral judgment M1 is a response to factor F1 in the situation under scrutiny.

Normative premise 1: Moral judgments that are responses to morally irrelevant factors of a given situation ought to be rejected because those factors would lead us astray;

Normative premise 2: F1 is morally irrelevant;

Normative conclusion: M1 ought to be disregarded from a moral viewpoint.

I think that the first normative premise is uncontroversial[137]. I grant it and I do not discuss it further. When I henceforth write "normative premise" of the AMIF, I refer to the second normative premise, unless explicitly noted. The AMIF is valid: if the premises are true, then the consequence follows. The AMIF (also known as the 'direct route') is a good way to derive normative conclusions from premises that are in part empirical. Since there are normative premises, this argument is compatible with Hume's Law[138]. However, the weak point of the argument is the Normative Premise 2, which is likely to be controversial. Hence, the argument will be sound for some people, and for others it will not. In what follows, I will mostly discuss Normative Premise 2 and its meaning.

I have shown in Ch. 2 that Greene and coworkers (2009) have found out that characteristically deontological judgments in scenarios such as *Footbridge* are driven by two factors: (1) a necessary means vs. foreseen side-effects distinction[139], and (2) personal force. Let us recall that Greene defined personal force (see § 2.1. above), in this way: "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." Greene actually thinks that both of these factors are morally irrelevant, because he thinks that the DDE does not hold. Yet he recognizes that the necessary means vs. foreseen side effect distinction is deemed morally important by many. In contrast, he thinks that personal force is considered as morally irrelevant by almost all people, both

[137] It is possible that morally irrelevant factors track morally relevant factors in a reliable way, though. I discuss this possibility below.

[138] It is impossible to derive a normative statement from a set of premises which are entirely descriptive. See Hume (1960/1739, 496). The literature on this topic is virtually unlimited. Hence, I prefer not to enter the topic at all.

[139] In what follows, I will interchangeably use "intention", "intentions", and "means vs. side-effect distinction" to indicate this factor.

laymen and professional philosophers. Hence, the AMIF can be re-written to target personal force, in the following way:

Empirical premise: Characteristically deontological judgments in *Footbridge* and similar dilemmas respond to the presence of personal force;

Normative premise 1: Moral judgments that are responses to morally irrelevant factors of a given situation ought to be discounted, because those factors would lead us astray;

Normative premise 2: Personal force is a morally irrelevant factor;

Normative conclusion: We ought to discount characteristically deontological judgments on *Footbridge* and similar cases.

This is the AMIF against personal force. This argument has a limited scope, since it applies to *Footbridge*-like dilemmas only. In spite of this, it raises some problems, so that in what follows I spell out a series of objections to this argument.

4.7. The AMIF against personal force: the interaction between personal force and intentions

My first objection to the AMIF on personal force is that the empirical premise does not take into account that, according to Greene et al. (2009, 370) (as quoted in § 2.2. above), System1 reacts to "representations such as 'goal-within-the-reach-of-muscle-force'." Moral judgments produced by System1 respond not to personal force *only*, but to a combination of intentions and personal force. There is a significant statistical interaction between the two. This means that personal force induces more deontological responses in participants only when coupled with the necessary means vs. foreseen side-effect distinction (cf. Greene et al 2009, fig. 4 at p. 368). In other words, actions that include *both* the use of personal force *and* harming others as a necessary means to reach one's goal elicit more deontological responses than other actions.

The moral relevance of intentions is well entrenched. The presence of some beliefs in the agent's mind, which can at least be linked to the means vs. side-effect distinction, is

described as morally relevant for the evaluations of many experimental participants (e.g. Cushman 2008; Young and Saxe 2008; Young et al. 2007). So it seems to be the case that laymen (often unconsciously) take into account the presence/absence of personal force and the presence/absence of intention when passing moral judgments. But of course it could be claimed that laymen are wrong in taking the necessary means vs. foreseen side-effect distinction as morally relevant. However, there are many moral philosophers, such as Quinn (1989b), that endorse the DDE. Therefore, the AMIF cannot be run against the means vs. side-effect distinction in the same way it is run against personal force. That distinction is not *prima facie* morally irrelevant. In contrast, there seem to be very few (or none) explicit defenders of personal force as a morally relevant factor. Nonetheless, what is more important for my objection is the influence of intentions on the AMIF against personal force. If Haidtian moral intuitions respond in *Footbridge*-like cases to both intention and personal force, i.e. to a factor that is morally irrelevant *and* to a factor that is maybe morally relevant, but is not *prima facie* irrelevant, can we say that these intuitions are off-track and ought to be rejected? This seems to be far from clear. Maybe tracking intention is so important for a correct moral assessment of a given scenario that we can tolerate tracking also personal force in some cases. Can we afford being led astray by the presence or absence of personal force, if this is the only way we have to track the means vs. side-effects distinction? It might be imagined that tracking intention correlates with tracking important moral facts and that, at least in some cases, tracking personal force correlates with tracking intention. This is a hypothetical case in which a *prima facie* morally irrelevant factor (personal force) tracks a *prima facie* morally relevant factor, i.e. intentions. What about the AMIF against personal force in this case? Of course it is again far from clear that it is empirically the case that these correlations hold and that personal force tracks intentions. Only further empirical work in experimental moral psychology will tell us whether tracking personal force is in some cases necessary to detect the means vs. side-effect distinction. Again, this is just a hypothesis so far. Be as it may, I can state that,

according to many moral philosophers (i.e. those that see intentions and agential beliefs as morally relevant), characteristically deontological judgments as described by Greene and colleagues (2009) do not track morally irrelevant factors *only*. They track a *mixture* of morally relevant and morally irrelevant factors, which makes their status and their credibility rather murky and confused. It is at least disputable that the AMIF against personal force can correctly run under these conditions.

It could be objected that, if one believes that p only because of (X&Y), where neither X and Y is sufficient for one's belief but both are necessary, then one should give up p if somebody demonstrates that X alone is not evidence for p. In other words, if one's intuition is a product of two necessary and jointly sufficient conditions, showing that either condition is not truth-conducive should be sufficient to demonstrate that the intuition is unreliable. In this case, if X is personal force and Y is the means v. side effect distinction, then showing that personal force is morally irrelevant ought to lead us to give up the deontological response to *Footbridge*. It is unclear whether this holds, both as a general principle and as in this specific case. If a judgment is a response to the conjunction of two factors, one relevant and one irrelevant, it is not clear whether the judgment must be discounted. The decision about this depends pretty much on the importance that is attributed to the relevant factor (in this case, the means v. side-effect distinction).

4.8. The AMIF against personal force: is personal force morally irrelevant?

It is also possible that personal force tracks a morally relevant factor in another way. Berker points out (2009, 324, footnote) that "it's one thing to say that whether one has committed a harm in an «up close and personal» manner is a morally irrelevant factor, and quite another thing to say that whether one has initiated a new threat that brings about serious bodily harm to another individual is a morally irrelevant factor." Of course, if A uses personal force on B, it can mean that A initiates a new threat that brings about serious bodily harm to B. So it is possible that there is a significant correlation between 'being an

instance of use of personal force' and 'initiating a new threat that brings about serious bodily harm' to somebody. Initiating such a new threat can safely be assumed to be morally relevant. If such a correlation was present, this would be a case of a *prima facie* morally irrelevant factor (personal force) tracking a morally relevant factor (initiating a new threat of the kind specified above). But this correspondence is by no means necessary: there are counterexamples. One may use personal force on a fellow human to perform an injection to cure a disease or to win a judo competition. There is plenty of uses of personal force that Westerners do not see as morally objectionable. However, whether this correlation holds is an empirical question that has not been addressed so far, at least to my knowledge. Moreover, it is possible to initiate a new threat that brings serious bodily harm even at a distance (e.g. by pulling the trigger of a firearm). None seems to think that the means by which the new threat is initiated is morally relevant. As Greene (forthcoming, § IV) puts it:

> Were a friend to call you from a footbridge seeking moral advice, would you say, 'Well, that depends… Will you be pushing or using a switch?' If questions such as this […] are not on your list of relevant moral questions, then you, too, should say "no" [when asked whether moral judgments ought to be sensitive to factors like personal force].

What morally matters is *that* a new threat that brings about serious bodily harm is initiated, not *how* this threat is brought about. Personal force seems to concern this latter aspect only. Contrary to Berker, I think that there is a general consensus about personal force as defined by Greene and colleagues in 2009 being a morally irrelevant factor. Of course deontologists will deny that characteristically deontological judgments respond to this factor, but this is an empirical issue, so that they must bring data in order to buttress this denial and to show that Greene and colleagues are wrong on this[140]. Summing up, I do not

[140] As W. Edwards Deming famously said: "In God we trust. All others must bring data."

think that personal force tracking the initiation of new threats of harm is an argument that makes any dent in Greene's armor.

4.9. The AMIF against personal force: judgments about moral relevance

My second and most important objection to Greene's neuromoral views comes from an analysis of judgments such as Normative Premise 2. I have granted Normative Premise 2 in the AMIF against personal force. However, before the argument can run, it ought to be examined what kind of judgment Normative Premise 2 is. That premise is a *judgment about the moral relevance of a factor*. The general form of such judgments is "X is a morally relevant factor." Several questions about these judgments need to be asked and I think that the AMIF is deeply problematic if they are not answered. For instance, is the belief "Personal force is a morally irrelevant factor" justified? If yes, how can it be justified? Does this judgment track anything morally significant? What factors drive this judgment, i.e. what are the psychological conditions that lead us to make it? What would the results of experimental inquiry on this judgment be? If it turned out that it depends on some Haidtian moral intuitions, how could it be shown that these intuitions are reliable and not unreliable?

I will carry out a preliminary analysis of judgments about moral relevance. This analysis is important to understand whether the AMIF succeeds, both against personal force and as a general argumentative strategy.

Let us start by making the safe assumption that card-carrying consequentialists like Singer and Greene[141] endorse this claim: "The number of people that incur harm in a given situation is morally relevant."[142] But what does this exactly mean? In order to explore this issue, it is useful to start by a suggestion of Hare's (1981, 63), who writes: "It is a mistake to suppose that we could first pick out the morally relevant features of a situation and only

[141] Also Unger (1996), that I mention below, is likely to endorse this claim.

[142] It is interesting to note that also this apparently uncontroversial claim has been contested (Taurek 1977), as I have written above. For a reasonable comment on Taurek's famous paper, see Parfit (1978).

then start asking what moral principles to apply to the situation. It is the principles which determine what is relevant." If a factor is morally relevant in a given moral framework, then there is some principle that appeals to that factor, and vice versa. Kumar and Campbell (2012, 318) claim that "a principle that something is (or is not) a relevant difference is justified because it makes sense of distinctions we intuitively make (or refuse to make)." As I understand it, this position is not significantly different from Hare's (1981), at least for my purposes, even though it involves a relationship of justification Hare does not mention. Hare actually believes that *prima facie* moral principles are established at the intuitive level of moral thinking and that claims about the moral relevance of factors are connected with those principles[143]. In Hare's opinion (1981, 63), then, talk about moral significance relates to the intuitive level only: "It is only when we come to intuitive thinking, guided by relatively general *prima facie* principles, that we need to be able to pick out the morally relevant features of situations, so as to leave out of consideration all the other features." Hence, in Hare's opinion, the moral relevance of factors has to do with the intuitive (and not the critical) level of moral thinking. As Hare put it, the critical level of moral thinking is the one that has to solve the controversies between distinct moral intuitions and "consists in making a choice under the constraints imposed by the logical properties of the moral concepts and by the non-moral facts, and by nothing else" (Hare 1981, 51). As a matter of fact, it is an utilitarian way of reasoning. The fact that judgments about moral relevance do *not* belong to the critical level, which is loosely analogous to Greene's System2, makes us wonder whether those judgments, that are so important for the AMIF, are reliable in Greene's view. Furthermore, Hare (1981, 100) writes that claims about moral relevance must be initially provided through a "guesswork," which seems to

[143] Hare thinks that *prima facie* principles like "One ought never to do an act which is F," where F is a certain property, are typical of the intuitive level of moral reasoning. Of course such principles can clash with each other and the critical level of moral thinking is engaged to solve these instances of conflict. The clash is usually caused by the fact that "one of the principles picks out certain features of the situation as relevant […] and the other picks out certain others" (Hare 1981, 52) Besides, *prima facie* moral principles "will have to be unspecific enough to cover a variety of situations all of which have certain salient features in common" (Hare 1981, 36). So the talk about 'salient features', i.e. factors of a situation, primarily belongs to the intuitive level and not to the critical level.

indicate that intuitions have a role in the elaboration of these judgments. Hare's claim is supported by the general epistemological principle that all judgments are influenced by the theory in which they are embedded. Judgments about moral relevance are made inside a given moral framework and this framework is defined by the general principles that are valid in it. Hence, these judgments importantly depend on the overall architecture of the moral context in which they are made.

Further light on judgments of the form "X is a morally irrelevant factor" comes from the considerations by Campbell and Kumar (2012) on moral inconsistency. They take up the dual process model, but in the weak form I mentioned in § 2.3.[144]. They start from the uncontroversial fact that humans do not like moral inconsistency. As I have hinted in § 3.1., suppose there are two cases or situations, S1 and S2 (which Campbell and Kumar sometimes refer to as 'target' and 'base' and which may be either real or hypothetical), and two moral reactions to these cases, M1 and M2. Then, if S1 and S2 are not separated by any morally relevant difference *and* M1 and M2 are very different from each other (i.e. approval vs. disapprobation), then moral inconsistency obtains. Moreover, the moral judge is usually more confident of her reaction to the 'base' case (say, M1) than of her reaction to the 'target' case, often a new case (say, M2). Moral inconsistency generates negative moral intuitions *à la* Haidt. One automatically feels moral disapprobation toward individuals, including oneself, when they exhibit moral inconsistency. Since moral inconsistency generates negative moral emotion and emotion is often taken to have an intrinsic motivational force[145], moral inconsistency motivates us to get rid of itself. How to do so? Normally, through System2, that carries out top-down regulation of emotion. More specifically, moral inconsistency is likely to be resolved by modifying the emotional response to S2, i.e. M2, because the moral judge is more certain of her M1 than of her M2. What I find particularly interesting to my purposes is that Campbell and Kumar claim that

[144] The characteristics of the two Systems are unchanged, but the connection with deontology and consequentialism is eliminated.

[145] This is questionable, but Campbell and Kumar seem to assume this and I do not want to discuss this point here. For more about this issue, see my (2012).

moral inconsistency is evaluated by both System1 and System2. System1 is seen as containing a series of norms that respond to particular features of cases[146]. The exact mechanism according to which the human mind deals with moral inconsistency is the following:

(1) In the presence of different moral reactions to a pair of cases, the moral judge identifies the salient differences between them – this is done through System2.

(2) These differences are fed as input to the moral judge's System1.

(3) If none of the norms that are present in System1 is activated, then System1 issues in a negative affective response.

The negative affective response flags one's moral responses to these two cases as inadequate. Norms may be thought as having a conditional form, e.g. "If X is Y, then X is morally wrong." A difference is structured in this way: "A is B, whereas C is non-B." If both B and non-B, i.e. both horns of the difference, trigger none of the norms present in System1, i.e. if both B and non-B are not Y (any of the Ys of the various norms), then System1 issues in an unpleasant emotional experience.

This may be clarified through an example. For instance, one could have a consequentialist response to *Switch* and a deontological response to *Remote Footbridge*, a *Footbridge* case in which the man with the backpack is let fall by using a remote control that opens a trapdoor under him. In this case, System2 identifies as salient the necessary means vs. foreseen side-effect difference: in *Switch* the death of the man is a foreseen side-effect, whereas in *Remote Footbridge* his death is a necessary means to save the five from doom. This difference is fed as input to System1. If none of the norms in System1 is triggered, then System1 generates a negative affective response and signals to other cognitive systems that there is something bad going on with these judgments. The moral judge will think that the two cases are sufficiently similar for the same type of moral response to be

[146] It is not important here to know whether these norms are something like the 'second-order modules' postulated by Haidt and Joseph (2007) and criticized by Suhler and Churchland (2011) or something else. Cf. § 3.1. above.

expected. In contrast, the moral judge has issued two very different responses, which puts her in a difficult situation.

A morally relevant difference between two cases is normally experienced as morally relevant in an emotional and motivational way. According to Campbell and Kumar, only System1 can generate this kind of experience, so that there cannot be correct judgments about moral relevance if System1 is not working properly. This creates what I find a very interesting phenomenon, that Campbell and Kumar dub as 'second-order moral dumbfounding'. Similarly to the original 'moral dumbfounding' in Haidt (2001), this phenomenon indicates a lack of justifications. In particular, people cannot justify *why* they find a factor either morally significant or insignificant. For example, a person could be asked: "Do you think that the number of people that die in this scenario is a morally relevant factor?" and could answer: "Of course it is!". Then she could be asked: "Why do you think so?" She would find it problematic to give an answer to that. The idea of second-order moral dumbfounding strengthens the aforementioned analogy posited by Hare (1981), according to which judgments about moral relevance and *prima facie* moral principles are equivalent, as they both elicit this phenomenon. A *prima facie* general principle such as "All cases of consensual incest are morally wrong" elicits dumbfounding and is difficult to justify, as we know from the empirical data reviewed in Ch. 3. As both judgments about relevance and *prima facie* general principles elicit moral dumbfounding, they are likely to arise from unconscious processing, i.e. from the activity of System1. This however does not guarantee that they are equivalent. The fact that they share this feature is just additional evidence to buttress the case of equivalence, but it is not conclusive evidence. It is Hare's argument for equivalence that does most of the explanatory work here. Be as it may, if both kinds of judgments arise from the activity of System1, then it is true that, as Campbell and Kumar (2012, 297) put it, inconsistency is *per se* "invisible to System2." Similarly, System2 is probably not involved in forming *prima facie* general principles such as the one concerning incest. System2 may "identify ways in which the

target situation and the base situation are dissimilar in morally relevant ways" and exert top-down regulation of emotion, but cannot by itself detect or deal with inconsistency. As I have written above in § 3.1., Campbell and Kumar do not marshal empirical evidence to support their descriptive hypothesis. Hence, their ideas just remain a from-the-armchair philosophical hypothesis. However, what would the consequences be for Greene's neuromoral claim if second-order moral dumbfounding actually existed and System1 were central to judgments about moral relevance? Could these judgments be considered 'reliable'? If so, under which conditions? The possibility that these important judgments are in turn based on cognitive systems that Greene takes to be blind and often unreliable is open. Greene owes answers to these questions if the AMIF on personal force is to run. I hope he will address these issues in his subsequent research. This concludes my preliminary analysis concerning judgments about moral relevance. This analysis was important because Normative Premise 2 of the AMIF against personal force is such a judgment.

To my purposes in discussing Greene's neuromoral theory, another important point that emerges from this analysis of judgments of moral relevance is that these judgments are moral in nature (since they are equivalent to moral principles) and as likely to be controversial as claims about the correctness of general moral principles or moral judgments about specific cases.

Summing up, I claim that judgments about moral relevance necessarily involve reference to a (sometimes implicit) moral principle. If a moral principle M1 that refers to factor F1 holds in a given moral framework, then F1 is morally relevant inside that framework. Vice versa, if F1 is morally relevant inside a given framework, then there will be some moral principle M1 in that moral framework that refers to F1. The number of victims of a situation is usually seen as morally relevant *because* most humans endorse a principle like "If the only difference between two alternative outcomes is that in the one outcome more people will suffer some harm than in the other, then select and strive for the outcome in

which the lesser number of people incur that specific harm." Hence, a factor is morally relevant for a judge when the judge already endorses a principle that takes that aspect of reality into account. The evaluation of moral relevance is not made in a moral vacuum. On the contrary, it expresses the morality one already endorses. So we must expect at the level of attribution of moral relevance some degree of disagreement, since we find disagreement at the level of moral principles. For instance, examine the judgment "Ethnic affiliation is a morally irrelevant factor" and the moral principle "It is morally wrong to judge differently two cases if the only difference between them is that in the one action the agent was an African American man and in the other action the agent was a Caucasian man." These two claims are equally controversial. Given that disagreement about some moral principles such as the DDA, the DDE, and many others is rampant and long-standing, it is exceedingly unlikely that the assessment concerning the moral relevance of factors such "being an action as opposed to an omission", "being a necessary means instead of a foreseen side-effect", and so on, will not stir considerable amounts of disagreement. Hence, moving from the level of judgments to the level of which factors of a given scenario count as relevant does not lead the debate to a significantly higher level of consensus, i.e. closer to solution.

An ambiguity ought to be dispelled here. When one says "X is morally relevant", one usually makes, as I have said, a normative point. For instance, Greene is claiming that there should be no moral principle that appeals to personal force. It would be wrong for humans to deem personal force morally relevant. This counts as a normative claim. However, the judgment "Personal force is morally irrelevant" could also be seen, in a context such as experimental moral psychology, as a descriptive judgment. This is known in moral philosophy as 'morality in inverted commas' (Hare 1952, 18-19), i.e. morality considered as a social phenomenon, from the descriptive point of view only. So, when one says "X is a morally relevant factor," one could mean (even though it is rare) that "People take X as a morally relevant factor." If Greene et al (2009) are correct, a descriptive claim

"Personal force is a morally relevant factor" could be true. However, this must not be confused with the more common, *normative interpretation* of a judgment about moral relevance. Only this latter interpretation matches with the kind of judgments that are required for Normative Premise 2 in the AMIF against personal force. So Greene is saying that from the descriptive point of view "Personal force is a morally relevant factor" is true, in the sense that it is implicitly endorsed by experimental participants through their behavior, and that, at the same time, "Personal force is a morally relevant factor" is false from the normative point of view, as this factor should not count in the decision of moral judges. This normative claim opens up interesting questions. What kind of resources can be marshaled to justify it? What kind of normativity is Greene appealing to? In order to go deeper into this issue, we have to examine again the meta-normativity problem. In what follows, I will avail myself of "X is a morally irrelevant factor" in the normative sense, unless explicitly noted.

4.10. The meta-normativity problem

When Greene says that improved knowledge of the machinery for moral judgments makes humans understand that some of their moral judgments about cases track morally irrelevant factors and that this in turn may be used to make "better" moral judgments, what is the meaning of this "better?" Although Greene is not entirely explicit on this, it is quite clear that he has a moral interpretation in mind. The analogy with capital juries he makes in (forthcoming) suggests that he intends the 'oughts' to be understood in moral terms. When one says that capital judgments ought not to be sensitive to race, the 'ought' is most plausibly interpreted as a moral one. It is *morally* wrong for a jury to make decisions about capital punishment on the basis of racist prejudices, as well as it is *morally* wrong for jurors to rely on their intuitions if such intuitions are (known to be) affected by unconscious racist biases, because racist biases seem to be morally irrelevant. As I have argued, saying that a factor is morally irrelevant seems equivalent (in the more common

interpretation; see the preceding §) to saying that there should be no moral principle that tracks that factor. It is a normative and not a descriptive claim. Hence, the decisions the factor affects are likely to be *morally wrong* decisions in which the moral judge has been led astray. And, surely, it is *morally* wrong to (knowingly) make *morally* wrong decisions. After all, *morally* wrong decisions are likely to lead to *morally* wrong actions[147]. Hence, Greene seems to solve the meta-normativity problem by choosing the moral horn and excluding the use of different forms of normativity, such as prudence. This means that, in Greene's view, the normative interpretation of the judgment "Personal force is a morally irrelevant factor" counts as a moral judgment. This fits well with my (and Hare's) idea that claims about morally relevant / irrelevant factors and claims about moral principles are equivalent. So, if we grant the normative premise of the AMIF, we face a conflict between moral judgments. On the one hand we have the common and widely held deontological response to *Footbridge*. On the other hand there is the commonly held view that personal force is a morally irrelevant factor. The conflict has two sides: a descriptive one and a normative one.

From the descriptive point of view, one could ask: "Would people choose to keep the judgment about moral relevance or the deontological response to *Footbridge*?"

From the normative point of view, one could ask: "Should people choose to keep the judgment about moral relevance or the deontological response to *Footbridge*?"

From the descriptive standpoint, we do not know how many of those who make the deontological judgment on this scenario (more or less 80% of the participants) would dump the deontological judgment in order not to be committed to the thesis that personal force is morally relevant. Greene seems to assume that most people would change their

[147] It could be argued that to be wrong about moral facts is morally wrong, and that this moral wrongness is supplementary relative to the wrongness of a moral action that originates from such erroneous moral knowledge. Alternatively, it could be argued that errors in moral knowledge are morally wrong just because they lead to morally wrong actions. I do not want to enter this interesting issue here, since what concerns me at the moment is to show that to follow a morally irrelevant factor is likely to be an instance of moral wrongness (as opposed to other kinds of wrongness).

minds in front of the AMIF, but this is not clear. We need empirical data to know whether the AMIF can (descriptively) be an effective tool of moral reform[148].

From the normative standpoint, the normative premise in the AMIF against personal force is a (general) moral judgment and the deontological response to *Footbridge* is another (particular[149]) moral judgment. There is a clash between two very widely endorsed moral judgments and Greene does not provide justifications for his (normative) claim that we *ought to* prefer one over the other but the fact that lots of people think that personal force is morally irrelevant. But this is a descriptive statement. It is not clear how it can *per se* justify a normative claim. Furthermore, we must descriptively take into account that also the deontological response to *Footbridge* is widely shared. This clearly shows that this kind of justification is insufficient. Greene owes his readers a better answer to the question as to *why* we *ought to* solve this conflict in the way he suggests.

Of course Greene could have gone prudential, so to speak. Greene's normative claim that people ought to dump the deontological response to *Footbridge* and to push the man down to his death instead could be interpreted prudentially, so that it is *prudentially* better if people sacrifice one life to save five. On this view, avoiding following the deontological response and adopting a more consequentialist way of thinking would be the best way people have – when faced with this specific case – to secure some important desired outcomes, such as the reduction of suffering. This is an interesting view and it seems capable to avoid many of the problems raised by the view discussed above, the one according to which the 'oughts' are moral ones. But this alternative view may well have to face other difficult problems. Only a thorough exploration (both psychological and philosophical) can tell us. The only thing I would like to point out here is that the truth of the claim that one ought (prudentially) to block one's deontological responses about these

[148] I commit here to no claim both about whether moral reform is desirable in general and about which kind of moral reform Westerners ought to strive for. I am completely agnostic about this. However, Singer and Greene (together with Unger, that I mention below) are not and do see themselves as moral reformers.

[149] I.e. about a case – it is not a principle.

cases and that one ought (prudentially) to avail oneself of a more consequentialist way of thinking is an empirical issue, and a difficult one to investigate.

4.11. An argument by Kumar and Campbell

Before moving to a more general discussion of the AMIF in the context of Greene's neuromoral theory, I would like to dispel an argument about this theory that in my opinion does not work. Kumar and Campbell think that the AMIF against personal force in the form I presented cannot run *even* if it is granted that the difference between physical harm inflicted through personal force and physical harm inflicted via other means *is*, indeed, morally irrelevant. I agree on this, but I do not agree with the motivations they adduce and I think that their point is no hindrance to the AMIF against personal force. Kumar and Campbell argue that Greene's AMIF on personal force works like that:

> [1] The deontological judgment about Footbridge is a response to personal harm.
>
> [2] The consequentialist judgment about Bystander [aka *Switch*] is a response to impersonal harm.
>
> [3] The difference between personal harm and impersonal harm is morally irrelevant. Therefore, [4] Either the deontological judgment about Footbridge or the consequentialist judgment about Bystander is unwarranted.[150] (Kumar and Campbell 2012, 317)

So Kumar and Campbell think that the argument made by Greene does not undermine the deontological response to *Footbridge*, but *just* the possibility of reconciling it with the consequentialist response to *Switch*. They are accusing Greene of concluding *too much* from his argument. But it is very likely that Greene would not grant [2]. He would say that the (empirically frequent) consequentialist response to *Switch* relies on the body count, and he would *make the normative assumption that the body count is morally relevant*, which is quite a safe assumption, as almost everybody except Taurek (1977) thinks that the numbers

[150] I modified the numbers of the propositions relative to the original text.

count, at least *ceteris paribus*. Greene is not making the point that the two responses conflict – this is rather obvious. He is claiming that one response tracks a morally relevant factor (i.e. body count) and that the other response tracks (together with intention) a morally irrelevant factor (i.e. the distinction between personal and impersonal violence, e.g. shooting a long-range rifle vs. punching in the head), so that the latter *ought to be rejected*[151]. He assumes what factors are relevant or irrelevant, as he thinks these claims are uncontroversial. As explained above, this is not so obvious.

4.12. The extended AMIF

This being said, Greene tries to extend the AMIF against personal force to other factors, in order to show that characteristically deontological judgments are unreliable under a broad series of circumstances. Greene aims indeed at a general debunking of deontology. In particular, for reasons I examine below, System2 must in Greene's opinion override System1 in all cases of conflict. This opens up some further issues. There are some cases that elicit a clash between deontological responses and consequentialist principles, but in which the contested factor is not personal force, but spatial distance. The most famous case is yielded by the two scenarios in Singer (1972). We feel compelled to help a dying child that is going to drown in a shallow pond (*Drowning Child* case), but we do not feel compelled to help a dying child that is starving in some distant village in troubled Bangladesh (*Envelope* case). We usually condemn those that do not help the drowning child, but we do not morally condemn those who fail to help the Bangladeshi child[152]. What drives this differential response?[153] Jay Musen, in a still unpublished Honors thesis in Greene's lab at Harvard, has experimentally shown that this differential evaluation is descriptively driven by spatial distance. People are actually influenced by this factor when

[151] Because Greene assumes Normative Premise 1 of the AMIF, above.
[152] Unger (1996) is a thoughtful elaboration starting from this dilemma.
[153] It is interesting to notice, as Campbell and Kumar (2012) correctly point out, that Singer's (1972) argument as to whether there is no relevant moral difference between *Drowning Child* and *Envelope* does not involve a utilitarian principle. It is an argument that is based on the seeming implausibility of considering distance as a legitimate moral principle.

making moral judgments and "Mere spatial distance is a morally relevant factor" is *descriptively* true. Greene, together with Singer and Unger (1996), deems this factor morally irrelevant *from the normative point of view*. Is it indeed morally irrelevant? In the personal force case, we could reach agreement (there is virtually no one that thinks that personal force as described by Greene is morally relevant) and we were left puzzling about the meaning and normative significance of this agreement. In this case things fare worse for the AMIF. There *are* moral philosophers that defend the moral relevance of distance. For example, Kamm (2007) spends two full chapters (ch. 11 & 12) to articulate how spatial distance influences justified moral judgments, i.e. the moral judgments we ought to make. The title of Kamm's book (2007) is *Intricate Ethics*. It is thus a foregone conclusion that her position is highly complex. In what follows I will just sketch the basic lines of her articulated position.

Kamm takes into account distance only *relative to the duty one has to help other humans who are in danger*. Hence, her discussion is clearly inspired by, and a response to, the views put forward by Singer (1972) and Unger (1996). Kamm's first claim is that *Drowning Child* and *Envelope* are not equalized. The two cases differ in many respects apart from distance. We cannot know for sure what drives participants' differential responses to that pair of cases. Hence, we need equalized cases, that differ on distance only. An example of equalized cases is the following:

> Near Alone Case: 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 Case: 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, all I must do is put the $500 I carry in my pocket into a machine that then triggers (via electric current) rescue machinery that will certainly scoop him out. (Kamm 2007, 348)

Kamm thinks that there is an intuitive difference between the two cases: we have stronger obligations in the former scenario than in the latter, which does not mean that we have no obligation in the latter. In particular, in the former case the agent can be required to pay more costs and to make more efforts to help the victim. For instance, it might be the case that morality demands from the agent to pay an even larger amount of money[154].

Then she argues for the moral relevance of distance. Against Singer and Unger, she writes that

> We cannot conclude that distance is never morally relevant simply by showing that one time or even sometimes it makes no difference to the strength of a duty in equalized cases. This is because a property that makes no moral difference in some equalized contexts may make a difference in other equalized contexts. By contrast, we can show that distance is morally relevant by showing that we think it matters morally sometimes—even one time—in equal contexts, even if it does not always make a moral difference. (Kamm 2007, 348)

After having established that distance morally matters, since it creates a moral difference in equal contexts as shown by responses to the Near Alone Case and the Far Alone Case, Kamm examines both the 'Standard view' and the 'Standard claim'.

The 'Standard View' is the idea that the problem of distance in morality is "whether we have a stronger duty to aid strangers who are physically near to us just because they are physically near than we have to aid strangers who are not physically near, all other things being equal" (Kamm 2007, 345). In her opinion, describing the problem of distance in morality in such a way that it involves reference to the distance between ourselves and victims only is misleading, since the problem may also pertain to the distance between means to help (e.g. a car) and victims, or to the one between ourselves and threats (e.g. a runaway trolley), or between means to help we own and threats, and so on.

[154] Kamm recognizes the difficulty to draw any clear-cut boundary relative to what we owe to each other in this sense.

The ‘Standard Claim’ derives from the ‘Standard View’. It is the idea that, if distance matters morally, we have a stronger duty to aid a near stranger than a far stranger, given their equal need. Kamm thinks that this is actually false. She maintains that the moral relevance of distance to helping does not conflict with the intuition that we have a strong obligation to help distant strangers. In her opinion Westerners have a strong intuition that, when the threat is near to the agent but the victim is far, the agent still has a strong duty to help. The agent ought to stop the nearby threat from hurting the faraway victim, e.g. by stopping a near killer from firing a rifle. Hence, if nearness is intuitively important, “this very fact may imply that we have strong obligations to aid distant strangers.” (Kamm 2007, 369).

The overall point which Kamm is driving at is that “we have greater obligations to take care of what is in the area near us (victims, threats, means)” (Kamm 2007, 370) and that “the responsibility to aid the near victim is stronger than the duty to stop near threats, and both obligations are stronger than the duty to rescue a victim’s near means or activate an agent’s distant means near a victim, other things being equal.” (Kamm 2007, 377).

Therefore, the problem of distance in morality ought to be understood in this alternative way: can we “justify our intuition that we have a greater responsibility to take care of what is going on in the area near us or near our (efficacious) means, whether this involves needy strangers, threats, or means belonging to strangers” (Kamm 2007, 376)? Kamm admits that intuitions are not self-justifying. They need to be connected with some principles that are unquestionably morally relevant in order to be justified. In order to justify the distance effects described above, Kamm maintains that there is a personal prerogative to give greater importance to one’s own interests and projects rather than giving equal weight to oneself and to others. In other words, we are morally allowed to be partial[155]. If an agent chooses to be partial, she gives importance to things out of proportion to the weight they

[155] Of course moral theorists such as Singer (1981) would not necessarily agree with this, even though they could accept partiality if it turned out to be a good way to maximize the furthering of the interests of sentient beings.

have from an impartial perspective. However, this option to be partial also creates a duty generated from the perspective on life from which the agent is acting. This duty is to take care of what is associated with the agent, for example, the area near her and her means. The moral import of distance therefore depends on adopting this partial stance in one's moral life. The person who has chosen impartiality could disregard distance effects. Finally, Kamm rejects Unger's (1996) debunking explanation according to which distance effects are due to salience. According to Kamm, it is true that in Near Alone Case and in Far Alone Case there is a difference in the emotional salience of victim's need, so that salience itself is much less in the far case than in the near. Nonetheless, she also thinks that, if we equalize these cases relative to salience, we still find a difference in Westerners' responses: "When the Near Alone and Far Alone cases also both have salient need, it is nearness and not salience that gives rise to our intuition that we have a strong obligation to help in the Near Alone Case" (Kamm 2007, 357).

I do not want to discuss whether Kamm is right about both this latter point against Unger and her previous point concerning the justification of distance effects through the personal prerogative to partiality. Whether she is right or not, Kamm constitutes a clear example of a contemporary moral philosopher who strongly asserts the moral significance of distance for humans' obligations to help fellow humans in need. As a consequence, it is far from clear whether spatial distance is a morally relevant factor or, on the contrary, a morally irrelevant factor. Therefore, we cannot run the AMIF for scenarios that look like *Envelope*. We usually pass judgments of mere moral acceptability for this behavior, at variance with the famous *Drowning Child*. But the normative premise that distance is a morally irrelevant factor (from the normative point of view) cannot be granted without evoking a substantial moral disagreement. Greene hopes to make deontology collapse by appealing to some claims about the moral relevance of factors that he deems widely accepted. However, he will persuade only those that do not hold moral principles that are connected with allegedly morally irrelevant factors. Greene's argument works, but only for those that grant

the normative premise. And it is not clear whether those who will grant the premise about the moral irrelevance of distance, thereby jettisoning the widely held deontological response to *Envelope*, are significantly more than those that endorsed the consequentialist solution to *Envelope* in the first place.

4.13. The deontological response to the extended AMIF

At this point, when confronted with an extended AMIF, the deontologist still has some tricks to play. She could easily say that the consequentialist response to *Footbridge fails to track* some factors that the deontologist sees as *morally relevant*, such as the separateness of persons, or the fact that human life is inviolable. In this latter case, she might argue, as Thomson (2008), that *Switch* and *Footbridge* ought to have the same solution, just as the utilitarian claims, but that this solution ought not to be that the agent should kill one to save five in both scenarios, but that she ought to let the five die in both scenarios because one morally relevant factor, i.e. that you cannot kill humans to save humans, overrides the harm minimization principle quoted above. Here not only the consequentialist theorist does not take into account an important factor, i.e. the inviolability of human life, but she also fails to understand that this factor trumps considerations concerning aggregate harm. Therefore, it is still very open to the deontologist to disagree with consequentialists such as Singer, Unger, and Greene about which factors are morally relevant and about the priority relationships between them. So Berker is perfectly correct when he writes: "In order to use the neuroscientific results and some philosophical theorizing to discount certain intuitive verdicts about trolley-like cases, Greene in effect needs to have already solved the trolley problem" (2009, 327). Finally, Berker (personal communication) argues that characteristically deontological and consequentialist responses to *Footbridge* are likely to respond to multiple factors and that the more adequate the description of these factors becomes, the more implausible their moral irrelevance will be. I agree with Berker on that. The idea that 'intention + personal force' is morally irrelevant is already disputable (cf. §

4.7.) If further factors are added, the claim of moral irrelevance could become even more implausible. Although empirical investigation only can tell what deontological judgments are exactly responding to in different cases, it is true that there seems to be a trade-off between the descriptive adequacy of empirical investigation and the claim that the factors tracked by deontological judgments are morally irrelevant.

Besides, Nichols and Mallon (2006) showed that the differential response to *Footbridge* and *Switch* is independent from the entities in danger being human beings. It appears with china cups too. Therefore, it may be conjectured that the personal force factor is not relevant to participants' responses because it is usually coupled with violence against animals. The causal story behind these responses seems to be more complex than we are inclined to think.

An important qualification is needed. Greene (forthcoming, § IV) specifies that "An empirically-driven normative argument is non-question-begging if the normative assumptions it requires are less interesting (i.e. less controversial) than its normative conclusion." So it is not required, in Greene's view, to find uncontroversial normative premises. Trading easy assumptions with trickier conclusions already constitutes moral progress, in his view. Empirical results can thus be used to develop particular traditions in moral thinking, such as for instance act utilitarianism. Traditions are identified through a set of normative assumptions. If empirical claims are added to these assumptions, you get more precise and controversial moral judgments that, however, will be valid inside the parochial circle of that specific tradition only. Those who do not grant the assumption will be utterly unimpressed by this kind of AMIF. Yet, Greene's idea of parceled moral progress, i.e. a moral progress that happens by empirical refinement of specific normative assumption inside separate moral traditions, clashes against the sweeping claims Greene himself makes. The idea in itself is interesting and even appealing. Surely it would be worth discussing. However, Greene himself does not seem to actually believe in it. Indeed, Greene does *not* argue that deontology is debunked for those who accept some normative

premises *only*. He thinks that deontology is debunked in all cases of conflict with consequentialism, or at least in all cases that are 'unfamiliar' in the sense described below, and for everybody. For any reasonable, adult human being, deontology is a bad idea because it is the rationalization of System1 intuitions[156], and those intuitions are often driven by morally irrelevant factors. This is Greene's view. If Greene wants to limit the scope of his empirical debunking of deontology to those who grant that personal force, distance, and inviolability of life are morally irrelevant, then one of the problems for his neuromoral theory is solved[157]. But in this case, philosophers such as Thomson and Kamm could simply ignore his debunking argument, since they deny the assumptions of his argument.

Another interesting point concerns the role science plays in Greene's neuromoral argument. Berker (2009) claims that empirical research makes no work in Greene's argument. I think Berker is wrong on this. Experimental moral psychology, as Greene (2010, 14) correctly points out, makes us understand what the factors are which some moral judgments are responding to. I do not want to enter the empirical issue whether Greene persuasively showed that deontological judgments on *Footbridge* respond to intention + personal force, since I am assuming *arguendo* Greene's descriptive theory here. What I want to say is that empirical science can *in principle* tell what are the factors that causally influence specific moral judgment about cases[158]. No matter whether Greene has succeeded or not in his empirical research, it is clear that experimental psychology can do this and, I add, it is the only kind of inquiry that can yield this result. It is not possible to do

[156] Insofar as it is a reflective cognitive operation, deontology needs System2 to be elaborated. However, its raw material comes directly from System1. Hence, Greene sees it as a direct expression of System1, even if System2 is a necessary condition for its existence. If humans had no System2, there would be no deontology. Greene's idea is that moral reasoning in deontology works like individual moral reasoning in the SIM: it mostly confabulates explanations for emotionally-loaded moral intuitions.

[157] Other problems that I described above, such as the one concerning the interaction between personal force and intentions, would stay, though.

[158] Kumar and Campbell (2012, 322) agree with me on this: "Empirical research can be of service here. The research cannot of course tell us whether a proposed difference is morally relevant. What it can tell us is which of the usually many differences between the cases is driving the divergent responses."

this from the armchair, even though some, like Unger (1996)[159] and Kamm (2007), have tried and try to do so. In contrast, I agree with Berker (2009) and Levy (2007, 305) that deciding which factors are morally relevant / irrelevant is an armchair operation that has nothing to do with the lab. Nonetheless, Greene (2003) never claimed that the opposite was true and always stuck to Hume's Law.

Still on this topic, Greene (forthcoming, § IV) reports an objection by Tim Scanlon, according to which "the work done by the science, while not insignificant, is morally insignificant. The science does not challenge anyone's values. Instead, it simply alerts us to an application of the values we already have." I think Scanlon is perfectly right on this. In order to reply, Greene puts forward an empirical debunking of a widespread moral intuition about consensual incest. The argument runs in the following way:

Empirical Premise 1: The inclination to condemn incest of all kinds is based on an emotional response whose evolutionary function is to avoid producing offspring with genetic diseases.

Empirical Premise 2: Vasectomy makes birth defects impossible, since it makes conception impossible.

Normative Premise: If humans condemn consensual incest with vasectomy due to an emotional response that evolved to prevent birth defects, then we ought not to condemn this kind of incest.

Normative Conclusion: We ought not to condemn consensual incest with vasectomy.[160]

But empirical science does not seem in this case to introduce any new value, or to change values. The normative premise is substantial and arguably many people would not grant it, even though only careful empirical work can tell us this. An open alternative, for instance, is to think that incest is wrong *per se*, because it offends God or some other deity, or

[159] Unger (1996) tried to carry out some informal surveys to test his intuitions about cases, but his work is not conducted according to Null Hypothesis Significance Testing, which is the standard statistical procedure in this case.

[160] De Lazari-Radek and Singer (2012, 21) put forward a very similar argument, even though their argument is more explicitly an instance of the EDA.

because it offends human dignity. Those who hold this position will not accept the Normative Premise above and therefore will not change their values. Hence, this argument challenges nobody's values, but those of people who accept the Normative Premise above (who are likely to already accept the Normative Conclusion, by the way).

This can suffice as far as the AMIF is concerned.

4.14. The indirect route

Greene also uses a second, more general way to attack deontology. The extent to which this broader, indirect route depends on the direct one (which was constituted by the AMIF against personal force, distance, and other factors) is unclear. At present I will consider it as independent from the considerations and doubts above. Greene claims that System1 and System2 show an efficiency vs. flexibility trade-off. On the one hand, System1 is less flexible and more efficient, i.e. fastest and computationally less expensive, than System2. On the other hand, System2 is more flexible and less efficient than System1, since it can evaluate scenarios and situations to which the individual has not been previously exposed. Greene then argues that "Automatic settings [i.e. System1] can function well only when they have been shaped by trial-and-error experience" (forthcoming, § V). This kind of experience can also not be personal experience. It can be the experience of some individuals that lived before a given human being and that have passed their experience down to her through either genetic or cultural transmission. A given individual has thus a functioning System1 relative to a given scenario S1 *only if* she has some experience of S1 through (1) genetic transmission, or (2) cultural transmission, or finally (3) learning from personal experience. The further step in Greene's normative argument for the indirect route is to define unfamiliar problems as "the ones with which we have inadequate evolutionary, cultural, or personal experience" (*ibidem*). Yet it is difficult to understand what scenarios or problems human beings are unfamiliar with. Hence, Greene provides two proxies. First, "moral problems that arise from recent cultural developments, most notably the rise of

modern technology and the intersection of disparate cultures, are especially likely to be unfamiliar." Secondly, moral disagreement which is not due to diverging opinions on non-moral facts is another tell-tale sign that a scenario or problem is unfamiliar. Therefore, Greene claims that we ought to rely more on System2 than on System1 when dealing with scenarios or problems that either elicit disagreement or derive from modern tech.

There are some issues relative to this line of argument.

First, most hypothetical scenarios used by Greene and other experimental moral psychologists are unfamiliar (relative to the above definition of this term). Westerners have undergone neither genetic nor cultural learning for *Switch*, *Footbridge*, or *Loop*, nor do they have any direct experience of similar cases. It does not happen often to have to choose between the survival of one person in danger and the survival of five people in danger. Hence, these cases are likely to lead System1 astray, if Greene is descriptively correct. So a problem arises. As I have written in Ch. 2, the empirical evidence for Greene's descriptive model mostly depends on experiments based on "trolleyology." But System1 is likely not to work well relative to these unfamiliar cases. Hence, during the experiments, participants will exhibit a defective functioning of System1, from the descriptive point of view. In other words, their System1 would not work as it works in most cases in real life, i.e. when confronted with familiar problems and cases. So, if Greene is correct, the view of System1 we derive from "trolleyology" is probably incorrect, again from the descriptive point of view, i.e. from the standpoint of experimental moral psychology. It might be the case that the dual-process model provides a deforming view of System1. System1 would look unreliable because it has been studied in cases for which it is not trained. Hence, it might also be the case that the dual-process model needs to be checked through "non-trolleyological" experiments, in particular to control that the picture of System1 proposed

by this model survives to experiments in which the participants are exposed to familiar situations[161]. In spite of this, I still assume *arguendo* that the dual-process model is correct.

Secondly, it is hard to see how moral disagreement could be a reliable indicator of 'unfamiliarity'. Consider for example those tendencies of System1 that were shaped by natural selection in the EEA and are universal in the human species. In the context of moral problems and scenarios that are relevantly different from those that our ancestors had to solve, such tendencies may well lead to what Greene would consider misleading emotion-based intuitions. But given that such tendencies are universal, everyone is likely to be misled *in the same way and in the same direction*. Unfamiliar moral dilemmas do not necessarily produce moral disagreement. Furthermore, it can be shown that moral disagreement can be present in problems that are perfectly familiar, for instance the death penalty, the moral admissibility of torture, the killing of tyrants that legitimately hold political power, or the justification of wars. Issues relative to torture and the death penalty were already present in work of Italian Enlightenment man Cesare Beccaria (1996/1764) in the late 18th Century. Focusing on the last moral problem in my list, in order to make a long story short, the existence of alleged 'just wars' was already discussed by Saint Thomas Aquinas in the 13th Century and by Immanuel Kant (1963/1795), so that Westerners seem to have quite a lot of cultural learning about this topic. This is not at all an unfamiliar problem (in Greene's sense), but there is much disagreement about it. People like Albert Schweitzer and Mohandas Karamchand Gandhi seemed to think that no war is just. People like former US President George Walker Bush seemed to think (and perhaps still think) that some wars are just. Hence, both of these claims hold: (1) there are cases of moral disagreement in which unfamiliarity plays no role; (2) there are cases of unfamiliarity that elicit no moral disagreement. As a consequence, moral disagreement and unfamiliarity are likely to be uncoupled.

[161] In a still unpublished Honors Thesis in Greene's Lab at Harvard, Katherine J. Ransohoff has probed participants with familiar cases, e.g. medical doctors with difficult medical choices. All of the scenarios were very high in ecological validity. Her results do not indict Greene's model. On the contrary, they confirm it. So it seems that the dual-process model passes quite well the test of high ecological validity experiments.

I doubt that Greene's criterion of unfamiliarity, in the sense defined above, is helpful. It is difficult to grasp what makes System1 on-track or off-track, since the issue seems to be in turn normatively loaded. It is so because to claim that a factor is morally relevant / irrelevant (in the normative sense) is to endorse a moral principle. System1 can then be on-track or off-track just relative to a set of moral principles. If the meta-normativity problem is solved by choosing the moral horn, then the assessment as to whether System1 responds in a way that is on-track or off-track when the individual is confronted with a given scenario or problem is relative to moral claims that the judge holds. In this case there is no way to decide whether and when System1 works *properly* in a moral vacuum, i.e. absent a set of moral principles. The meaning of 'properly' depends on those principles and, if they are nowhere to be found, that meaning is void. Lastly, empirical science can do no work in establishing which factors are morally relevant or irrelevant. Happily enough, Greene (2010) agrees with me on this latter point. But unfortunately Greene does not realize that to endorse the latter point also means that empirical science can do no work in establishing when System1 works *properly*.

4.15. Greene's argument against the DDE

A final part of Greene neuromoral argument is the following: an improved understanding of the machinery for moral judgments may cause humans to make "better" moral judgments because it can show that some moral principles are not valid. Hence, improved knowledge in moral psychology helps us not follow "bad" moral principles, i.e. avoid false moral beliefs, beliefs which usually correlate with carrying out morally wrong actions. In particular, Greene (forthcoming, § VI) runs an argument against the DDE[162]. In Greene's opinion, System2 can be used in two ways in moral philosophy: either as a tool to

[162] Another argument to the same effect is present in the draft of Greene's forthcoming book I was able to read. However, since Greene asked me not to quote from his draft, as it is still a work-in-progress, I refer to the (in my humble opinion) weaker argument that is present in Greene (forthcoming). It is an instance of straw-manning, but I was forced to carry it out. The argument against the DDE that is present in the draft of the book I was able to read does not concern RE and is based much more on experimental moral psychology than the one present in (forthcoming).

rationalize given moral intuitions, or as a tool to find out rational moral principles from which judgments about cases are derived in a deductive way. In Greene's opinion, deontology makes use of System2 in the former way, whereas consequentialism makes use of System2 in the latter way. This latter operation of System2 is called 'true reasoning' and it is independent of previous moral intuitions about cases. In contrast, it spots moral axioms that are similar to Sidgwick's philosophical intuitions. Since the consequences of these 'philosophical' intuitions can conflict with well-established Haidtian moral intuitions about cases, true reasoning involves a lot of "bullet-biting", i.e. rejecting those intuitions so that the principle can prevail. In contrast, rationalization chases intuitions and rejects them only at the margins, when RE dynamics push the theorist to do so in order to enforce consistency. At a maximum, deontologists can reject consequentialist judgments for the sake of consistency with deontological principles, an operation that Greene calls 'biting rubber bullets'. Greene claims that a principle such as the DDE derives its justificatory power from moral intuitions only, i.e. it is the result of the rationalization of unreliable Haidtian intuitions. In other words, the justificatory power of the DDE lies in intuitions only and provides no independent justificatory force. In Greene's opinion "the DDE was codified because it was observed that certain intuitive patterns in moral judgment could be summarized by a set of principles now known as the DDE" (*ibidem*). As evidence of this, Greene quotes the fact that the DDE gets discarded when it does not track Haidtian moral intuitions, such as in the *Loop* (also known as *Ned*) case, relative to which roughly half of the experimental participants accept to turn the trolley even though in that scenario the death of the one is a necessary means for reaching the end of saving the five.

There are some problems with this argument, just like with the previous ones.

First, it is not clear whether Greene's empirical investigation has persuasively shown that reasoning plays this kind of double role in the two normative theories under examination. There seem to be no empirical data that buttress this double-function interpretation of brain

areas such as the DLPFC and the ACC. However, I want to be charitable and to concede *arguendo* this empirical point.

Secondly, Greene claims that the DDE derives its justification from Haidtian moral intuitions. But the DDE is usually a part of deontological ethical theories and many of these theories justify moral claims through RE and variants thereof. So it is incorrect to claim that the DDE derives its justification from moral intuitions *only*, especially if WRE is considered. Under a WRE framework, the DDE would be justified because it is a part of an equilibrated set of CMJs, moral principles, and background theories. Hence, it would derive its justificatory force from all other components of the final state of WRE (final set of CMJs, final set of principles, and final set of background theories) and from the relationships between these items[163].

Besides, as I pointed out above, it cannot be taken for granted that there is a strong link between System1 and the various forms of RE, since RE starts from CMJs and not from moral intuitions *à la* Haidt. So Greene's analysis of the workings of System2 as a tool to rationalize and systematize the output of System1 in deontology tells us little about the goodness of RE and (especially) WRE as a method for justifying moral claims. Greene has not shown his readers so far that WRE is an unsound method to justify moral claims, provided that he actually wants to show this. However, he needs WRE to fall in order to show that the link between moral intuitions and the DDE is problematic. In other words, since in many deontological frameworks the justification of a principle such as the DDE depends on the workings of (one of the multifarious variants of) RE, Greene has to show that RE is a "bad" method to justify moral judgments if he wants to show that the DDE is unjustified. Alternatively, he has to show that no final state of equilibrium[164] including the

[163] It is unclear whether Greene wants to attack RE as a method of moral justification. On the one hand, he writes (forthcoming, § VI) that "Ethicists need to worry about their intuitions, and not just the ones that they're willing to dump in order to save the ones they really want to keep." This seems to be a critical description of a form of RE. On the other hand, he advocates a "double-wide reflective equilibrium" (*ibidem*) that includes moral intuitions, moral principles, background theories, and facts about moral psychology. Greene's views about moral justification are murky, if we take into account what he has published so far. I hope he will make things clearer in his forthcoming book.

[164] The state in which we are when we have stopped going 'back and forth' in RE.

DDE is possible starting from plausible CMJs, other moral principles, and background theories, i.e. that the DDE is excluded from any plausible final state of equilibrium. But this would be exceedingly complex to show and, luckily enough, Greene does not even attempt to do so. As Greene does not provide any convincing arguments to show that the multifarious variants of RE are wrong as methods of moral justification and does not take into account that moral intuitions *à la* Haidt are very unlikely to enter RE or WRE, his attack against the DDE will be properly assessed only when he makes his position on RE more explicit. So far, the argument concerning RE and the DDE is unfortunately not persuasive.

4.16. Levy's argument against the DDE

A more interesting attempt to debunk the DDE is the one made by Levy (2011). I examine it here because it is interesting to contrast it with the analogous endeavor by Greene. Levy tries to show that the DDE fails to provide any justificatory force because the attribution of intentionality, that is paramount according to the DDE to decide whether an action is morally acceptable or not, is normatively loaded. Levy is making reference to the famous Knobe (2003a, 2003b, 2006) effect[165]. Experimental philosopher Joshua Knobe has shown that if an effect E1 of action A1 is seen as positive by a judge and there are reasons to think that E1 is not the main end of A1, E1 is likely not to be considered as intended. On the contrary, and surprisingly, if E1 is seen as negative but the same reasons obtain, then E1 is more likely to be considered as intended. The effect is now well-known and a recent paper by Pinillos and colleagues (2011) shows that this effect is an emotional bias that tends to disappear in very reflective experimental participants. Hence, Levy's argument seems to be the following:

[165] This effect was actually first identified by Alicke (2000).

Empirical Premise 1: Attributions of intentionality of outcomes depend on substantive ethical claims held by the attributer – namely judgments as to whether the outcome is morally good or bad.

Empirical Premise 2: The DDE is regularly used to distinguish between morally acceptable and morally inacceptable actions.

Logical Premise: The DDE depends on attributions of intentionality.

Normative Premise: Principles that depend on substantive ethical claims cannot be used to discriminate between morally acceptable and inacceptable actions, since they simply re-instate and confirm the normative claims on which they are based[166].

Normative Conclusion: The DDE ought not to be used to distinguish between morally acceptable and morally inacceptable actions.

I grant the second Empirical Premise, the Logical Premise, and the Normative Premise. I have serious doubts about the first Empirical Premise, though, if this is interpreted without qualifications, as the argument seems to require. The problem is the same as in Walter Sinnott-Armstrong's skeptical claims (cf. § 3.6.): if we know that some instances of a phenomenon (e.g. attribution of intentionality, moral intuitions) are biased, are we then authorized to consider all instances of that phenomenon as biased? In particular, what is the threshold that must be overcome in order to allow generalization? Of course there will be a problem of vagueness of the boundary[167] and it will be impossible to draw any non-arbitrary cutoff. In particular, is the Knobe effect widespread enough to warrant the claim that all attributions of intentionality for morally relevant actions are biased and hence depending on substantive ethical claims? And is the effect powerful enough in terms of effect size to justify this claim? Or is the influence of the Knobe effect rare enough and weak enough to justify the claim that most attributions of intentionality humans carry out

[166] As Levy (2011, 6) puts it, "In what follows, I argue that these intuitions are sensitive to moral considerations in a way that makes appeal to them question-begging. It is question-begging because agents' preexisting moral views influence the application of the doctrine in such a manner that it generates the appropriate output."

[167] I.e. it will be difficult to decide whether the cases are enough or not to generalize if their number gets close to the threshold that has been posited.

are reliable, so that Empirical Premise 1 above cannot be granted? Finally, are there plausible alternative explanations of Knobe's empirical results?

Knobe (2003a, 2003b, 2006) describes a very strong effect: attribution of intentionality moves from 23% for a positive outcome to 82% for a negative outcome in the 2006 paper and from 30% for a positive outcome to 77% for a negative outcome in the second experiment of the 2003a paper. Therefore, it may be conceded that this effect is sizable. Furthermore, the body of evidence marshaled by Knobe ranges over quite different cases. They are not just variations on a same theme. Of course, interpretation of these results is not obvious. Adams and Steadman (2004) suggest that pragmatics could be involved. The participants could not want to say that a negative outcome is unintended because this would pragmatically imply that they condone the outcome and do not see it as morally bad. Since they do not want to revise their moral judgment and do not want to look as though they were shying away from it, they do not say that the outcome is unintentional even though their capacity of assessing intentions has established so. The experiments carried out by Knobe (2004) quite persuasively show that this alternative explanation is unjustified, but at the same time they show that there is a big difference, from the experimental point of view, between 'having the intention of carrying out A1 [and doing it]' and 'to perform A1 intentionally'. Subjects are much more likely to say that a negative outcome has been brought about intentionally than to say that the agent had the intention of bringing that outcome about. This is not easy to accommodate into Knobe's explanation of the effect, namely that attribution of intention is a multi-purpose tool that features the attribution of blame and praise as one of its main functions. For some reason, subjects seem to treat "to have an intention to do X" and "to do X intentionally" differently. Only further empirical work can solve this issue and properly answer the questions above. It is going to be very complex and hard empirical work. Attribution of intentionality ought to be measured in real life and in a longitudinal way to assess real-world frequency of the bias. Lots of different side-effects should be taken into account, even though Knobe has

already done much. Be as it may, Levy's argument is more persuasive than Greene's in (forthcoming), because it would work well if it could start from the right kind of empirical premises. Experimental moral psychology is unfortunately not there yet and I do not know whether it will ever warrant the empirical premises that are needed for this interesting argument to run. Nonetheless, I agree with Levy that "the data reviewed here presents a powerful challenge to a time-honored philosophical distinction" (2011, 8).

4.17. The *ad hominem* argument against the utilitarian theorist

Greene's position is explicitly utilitarian. A version of the "*ad hominem* argument" (henceforth AHA) has been used to attack the political conservative and could be used to attack the utilitarian. The AHA derives normative consequences from empirical facts. The AHA against the utilitarian counts as a neuromoral theory because it claims that improved knowledge in experimental moral psychology, namely that people having certain negative traits tend to carry out certain types of moral judgments, leads humans to reject those judgments as untrustworthy. As a consequence of rejecting those judgments, humans would make *better* moral judgments as a result of increased empirical knowledge. Greene does not put forth an AHA. On the contrary, an AHA, which is a neuromoral argument, could used against him. In this section I assess the AHA and I establish whether it is significant for Greene's normative position. It is unclear whether *anybody* is presently trying to attack the utilitarian on those grounds, but it is at least possible to do so. Furthermore, a recent paper has gone very close to doing so. This paper is Bartels and Pizarro (2011). Pizarro and Bartels have measured in a sizable amount of undergrads three traits: psychopathy, Machiavellianism[168], and the tendency to perceive live as meaningless. Then they administered 14 dilemmas that are similar to *Footbridge* and asked for moral

[168] Machiavellianism is "the degree to which people are cynical, emotionally detached from others, and manipulative." (Bartels and Pizarro 2011, 156)

acceptability of the consequentialist response[169]. As expected, just a few people gave consequentialist responses, but people who scored high for these three traits were significantly "more utilitarian." Please notice that all of the subjects were non-clinical. This result parallels the increased utilitarian responses shown by clinical groups such as low-anxiety psychopaths (Koenigs et al. 2011) and VMPFC patients[170] (Ciaramelli et al. 2007; Koenigs et al. 2007). It must also be noticed that Westerners currently tend to consider the three traits under consideration as negative. Westerners would like people to be neither psychopathic nor Machiavellian. Nor would they like that people perceive life as meaningless. Each Westerner holds a different picture of the ideal of the good life, but just a tiny minority of them would include these traits into the picture. From these results Bartels and Pizarro draw a methodological claim: psychologists ought not to use consequentialist standards as definitional of rational behavior in empirical studies. This is therefore another chapter of the 'rationality wars'[171]. Bartels and Pizarro argue that a behavior that is linked to disreputable personality traits ought not to be considered as a standard of rationality. Since utility-based standards are linked to these traits, Bartels and Pizarro implicitly question the position by Baron and Kahneman and seem to move closer to the views of Gigerenzer and followers. In their words,

> we should be wary of favoring a method that equates the quality of moral judgment with responses that are endorsed primarily by individuals who are likely perceived as less moral (because they possess traits like callousness and manipulativeness). In other words, adopting such a method can lead to the counterintuitive inference that ''correct'' moral judgments are most likely to be made by the individuals least likely to possess the character traits generally perceived as moral (Bartels and Pizarro, 2011, 158).

[169] Bartels and Pizarro (2011, 157) specify in a footnote that they re-run the statistical analysis of their data by only using the 7 dilemmas certified as pitting utilitarianism against deontology by Kahane and Shackel (2008). The re-analysis confirmed their results.
[170] For discussion about these patients, see § 2.2.
[171] See § 3.4.

However, Bartels and Pizarro do not claim that this shows consequentialism, or utilitarianism more specifically, to be unwarranted as a normative ethical theory, since they correctly acknowledge that "the characteristics of a theory's proponents cannot determine its normative status." What they argue against is "the validity of using these measures [i.e. utilitarian criteria] as a metric for optimal moral judgment in everyday life." But what is exactly the difference between a normative benchmark in experimental moral psychology, and a normative ethical theory about the moral good? I think that there is some difference, so that I doubt that Bartels and Pizarro are actually attacking the utilitarian[172]. A standard for rationality can be relative to an experimental setting only. For instance, suppose that an experimenter wants to measure the responses of participants to some morally relevant cases. Suppose further that, in the human group which the participants come from, there are some widespread moral intuitions, and that the researcher wants to compare the responses to a standard that is based on those intuitions. So she wants to check whether the responses from her sample comply with the moral intuitions that are present in the overall population. She could say that responses that conform to the population benchmark are 'rational' and those that are not conform are 'irrational'. However, she can find this standard completely misconceived in everyday life. She might *not* endorse and recommend the benchmark outside the lab. Hence, it might be the case that a psychological benchmark has no binding force and no motivational clout for the person who makes use of it. The researcher may avail herself of the norms of her society in inverted commas, i.e. simply using them as a yardstick to check whether a sample is similar to the population, but without taking them in earnest[173]. Hence, there is no necessary correspondence between claims about rationality in experimental moral psychology and claims about the moral good. Nonetheless, this correspondence is not impossible. For instance, when people like Baron (1994) defend a consequentialist framework to distinguish between rational and irrational decisions in experimental moral psychology, they seem to subscribe to a

[172] Besides, their attack would be very weak, so that it is also charitable to read them in this way.
[173] I owe this point to Matteo Mameli.

consequentialist ethical theory. One possible interpretation of what Bartels and Pizarro are saying is therefore that they are disguising their substantive ethical claims under the gowns of methodological wisdom[174]. But it cannot be shown that this is necessarily the case. Since they explicitly write that "the characteristics of a theory's proponents cannot determine its normative status", it is more plausible to interpret them as *not* running an AHA against the utilitarian. Nonetheless, it is interesting to examine an AHA against the utilitarian even though Bartels and Pizarro are actually *not* putting it forth.

This kind of AHA would go like this:

Normative Premise 1: If some people possess some psychological traits (psychopathy, Narcissism, Machiavellianism, tendency to see life as meaningless, etc) above a certain threshold, they have bad characters and they are bad people.

Normative Premise 2: If bad people endorse some moral claims significantly more than the rest of the population, that moral claim is unwarranted.

Empirical Premise: Utilitarian judgments about *Footbridge*-like dilemmas are made significantly more often by low-anxiety psychopaths and non-clinical individuals that are high on psychopathy, Narcissism, Machiavellianism, etc.

Normative Conclusion: Utilitarian judgments about *Footbridge*-like dilemmas ought to be deemed unwarranted.

A critic could object that nobody has actually put this argument forward: this is a blatant instance of straw-manning. As I have written above, it is correct that it has not been put forward (it is not correct that I am straw-manning, as I am attributing the argument to nobody). However, this point has been suggested. Levy (2007, 297) wrote that, due to the empirical results on VMPFC patients, utilitarian responses ought to be rejected as suspect. Furthermore, Marcus Arvan (2011, 2012) has run a very similar argument against the political conservative, especially against the conservative in social matters such as gay marriage, gun control, restriction of individual liberties to fight terrorist groups, death

[174] I thank Anna Pacholczyk for pointing this out to me.

penalty, and so on. After having replicated Bartels's and Pizarro's result that utilitarian responses to *Footbridge* correlate with Machiavellianism and psychopathy, Arvan found very significant, quite strong (in terms of Pearson's r) correlations with the Dark Triad[175] (Narcissism, Machiavellianism, and psychopathy) and typical conservative judgments in social matters. Then he explicitly runs the AHA against the conservative, quoting virtue-ethicist Hursthouse (2002, 28), who argued that we should understand morally right actions in terms of what the virtuous person would choose. He illustrates his point with this example:

> Suppose we knew (A) that people like Hitler are not morally virtuous individuals (a safe assumption), and we empirically demonstrated that (B) it is typically people like Hitler (i.e. "counteractive narcissists") who find anti-Semitic or racist moral views attractive. If both of these things were the case and Hursthouse's definition of right action is correct, then we would have strong inductive grounds for rejecting anti-Semitic and racist views on the basis of the personality traits to which they are empirically linked. (Arvan 2011, 9)

This argument is similar to the AHA against the utilitarian I reconstructed above. However, Arvan adds a further part to his AHA against the political conservative. This addendum usefully includes the notion of a trait threshold[176]:

> In order to determine which levels of the Dark Triad are morally bad, we should seek to determine which levels of those traits correlate with higher levels of behaviors that are widely or universally considered to be morally bad – for example, criminal activity. […] If particular levels of the Dark Triad are […] found to correlate significantly with the kinds of moral misbehaviors responders self-reported, we would then have real empirical evidence that those levels of the Dark Triad are related to morally bad behavior. What we could then do is test the data from my studies to see whether the morally bad levels of the Dark Triad correlate

[175] Please notice the normatively-loaded expression Arvan uses.
[176] As usual, the insertion of the threshold creates the familiar problems of vagueness and arbitrariness of the cutoff. There will be cases that fall in the 'grey area' around the threshold and the decision concerning the position of the threshold itself will have some degree of arbitrariness. Though these are interesting issues, I will not discuss them here.

> significantly with conservative or liberal judgments on the social issues […] If we did find that morally bad levels of the Dark Triad correlate with conservative (or alternately, liberal) views on particular issues […] then we would have strong correlational evidence that those (conservative or liberal) judgments are related to morally bad behavior. (Arvan 2012, 11)

Bad traits become "bad enough" to qualify the person as "bad" *only if* there is a significant correlation between levels of the traits that are above the threshold and some uncontroversial form of moral evil, such as violent crime (e.g. torture, rape, and murder of innocent toddlers). In the Hitler example above, this would mean that the "counteractive narcissist" is a bad person only if her score in Narcissism is high enough to significantly correlate with some uncontroversial form of moral evil, as specified above. In spite of this important addition, this argument is problematic. In what follows I will discuss together the AHA against conservatism and the one against utilitarianism, since their overall structure is similar. I am more interested in the overall form of the AHA than in its specific instances.

First, the notion of "people like Hitler" and "bad people" in my Normative Premise 1 seems to be problematic, even adding the threshold and the correlation with uncontroversial forms of moral evil. Doris (2002) has persuasively argued that results in social psychology have shown that people's behavior is influenced by the context much more than it was previously thought. The 'fundamental attribution error', according to which Westerners tend to attribute behavior to features internal to the agent rather than to contextual factors, is widespread, although contextual factors actually tend to be causally much more important than personality traits. The situation in which an individual is put exerts a powerful influence on her behavior, as shown by the well-known experiments by Milgram (1963), Zimbardo (2007), Latané and Darley (1970)[177]. So the concept of

[177] Milgram's and Zimbardo's experiments both show that people are prone to carry out actions that are normally regarded as immoral if authorized to do so by an authority (no pun intended) figure, such as a researcher. Latané's and Darley's experiments are about the so-called bystander effect: people help a person

character needs to be empirically validated. Contrary to provide a basis on which to build neuromoral arguments, the concept of character seems to be called into doubt by results in experimental psychology. Those who want to run the AHA either against the conservative or against the utilitarian ought, at the minimum, to provide a defense of the concept of character from the situationist challenge put forth by Doris.

Secondly, it cannot be taken for granted that traits like Machiavellianism and Narcissism are negative. There is a widespread moral intuition that they are morally bad, but of course this intuition could turn out to be unwarranted, just as intuitions against consensual incest with vasectomy seem to be unwarranted to many consequentialists. Many utilitarians spend most of their time trying to show that well-entrenched and cherished moral intuitions (*à la* Haidt) are unwarranted and should be thrown into history's dustbin. Hence, they are unlikely to buy moral intuitions at face value, including the one concerning these traits. Machiavellianism could be morally bad because it could correlate with a tendency to be deceptive, even though I know of no empirical study that has shown such a correlation. Narcissism could correlate with a tendency to ride roughshod over other human beings' interests, rights, and sensibilities. However, from the point of view of the consequentialist theorist what is morally bad is the act that causes a reduction in aggregate well-being or a setback of the interests of a sentient being, to use (roughly) Singer's wording. As I have shown, Arvan (2012) correctly builds his argument in such a way that a connection between the 'Dark triad' and "behaviors that are widely or universally considered to be morally bad" is required for the argument to run. Therefore, this latter version of his argument is more likely to be palatable to the consequentialist theorist than the previous (2011) one, in which this point was overlooked. The connection between psychopathy and violent crime is well-known, but I think reliable empirical material[178] must be marshaled to show that Machiavellianism, Narcissism, and the tendency to perceive life as meaningless

in need if alone, but do not help a person in the same state of need if somebody else can help too. The result is that the victim is often better off when she encounters a lonely potential helper than a whole group of them.

[178] And not some funny anecdotes on Hitler, Hobbes, and Rousseau, cf. Arvan (2011, 1-2)

are actually connected with behaviors that even consequentialists deem morally bad. This is an empirical issue and so it must be dealt with using the methods of experimental science. Arvan also suggested what kind of experiment to perform[179]. I hope that he (or others) will carry out this important empirical investigation. In the absence of this, the argument seriously risks begging the question against the consequentialist theorist.

Thirdly, even in presence of an established correlation between some psychological traits and some morally bad activity, this correlation could still be devoid of moral import. It is not clear whether the existence of hypothetical correlations between high Machiavellianism and lying and between high Machiavellianism and a consequentialist response to *Footbridge* tells us anything morally significant about the latter judgment. Lying is *prima facie* morally wrong: this is a safe assumption. But it could be objected that correlations are no conclusive evidence for causation and that as a consequence Machiavellianism is not problematic. Still, let us make the case of the upholder of the AHA stronger. Let us suppose *arguendo* that those correlations are causal links, i.e. that high Machiavellians lie significantly more frequently than others *because* they are emotionally blunted and that they pass that moral judgment *because* they are emotionally blunted. Does this undermine the judgment? Not necessarily. The judgment could have other causal histories in other populations. For instance, people with high need for cognition (Bartels 2008) tend to respond to *Footbridge* in an utilitarian way significantly more often than others. Hence, there is another explanation, different from emotional blunting, for consequentialist responses to *Footbridge*: extensive use of inferential reasoning can lead to that response. I stress that the explanation through emotional blunting is not a debunking explanation, because it is a partial explanation only. It solely holds for a subgroup of the population, namely those who score high in Machiavellianism, psychopathy, and meaninglessness of life. Utterly different psychological processes might

[179] Arvan put forward to use "the Comprehensive Misconduct Inventory, a 50-item survey which has participants self-report a wide variety of misbehaviors including criminal behavior, driving misconduct (e.g. "road rage"), bullying, alcohol and drug abuse, and aggression towards persons and structures of authority." (Arvan 2012, 11)

bring about those judgments in other groups. Perhaps those groups hold those judgments because they have carried out a brief cost-benefit analysis and have followed the dictates of the 'body count'. This third point is, I think, the reason why Greene (forthcoming, § V) boldly bites the bullet when faced with the AHA. He writes that people ought to rely more on System2 than on System1 when faced with unfamiliar problems. This implies, he argues, that clinical populations that happen to use System2 more than System1 across the board will behave *morally better* than 'normal' people when faced with unfamiliar problems. This is a simple contingency, that detracts no merit, in his opinion, from the claim that in unfamiliar cases we ought to use System2. I cannot but agree with Greene on this point.

Contrary to other arguments examined so far, I doubt the AHA will ever work, because the concept of character seems to be problematic on experimental grounds and because, even if it was possible to show that the correlations found by Bartels and Pizarro for the utilitarian and by Arvan for the political conservative are causal links, this would not amount to a debunking explanation for utilitarian and conservative judgments. Other avenues of justification for these judgments seem to be open. The fact that some sub-populations, no matter whether clinical or non-clinical, carry these judgments out because members of these populations are emotionally blunted, or callous, or cynical, or detached from human life, does not exclude that these judgments could be perfectly correct and justifiable. When the utilitarian is concerned, it is already known that very reflective and rational people tend to endorse utilitarian claims significantly more than others, so that an alternative causal explanation is at hand. Hence, Greene's neuromoral claims, that already have to deal with a broad and multifarious gamut of problems, do not have to add the AHA to the list.

Conclusion

In this thesis I have explored the complex relationship between normative ethics and experimental moral psychology by examining one widely discussed attempt to derive normative conclusions from empirical facts, i.e. Joshua Greene's neuromoral theory. According to Greene, an improved understanding of the machinery for moral judgments leads us to become wary of deontological tenets, especially when these ethical claims concern unfamiliar problems and conflict with competing consequentialist claims. Greene grounds this claim on a strong dual-process account of the machinery for moral judgments, that couples the fast and automatic System1 with deontology and the slow, conscious, and flexible System2 with consequentialism. In what precedes I have shown that:

(1) the empirical evidence available so far for the strong version of the dual-process model proposed by Greene is not conclusive;

(2) there are alternative accounts of the functioning of the machinery for moral judgments that cannot be ruled out by current empirical evidence;

(3) there are descriptive views (other than Greene's) that try to derive normative consequences from their descriptive content (e.g. Gigerenzer's) – in other words, there are other neuromoral theories than Greene's;

(4) these theories share with Greene's the meta-normativity problem, i.e. they do not clarify in what sense the advancement of empirical knowledge can bring about *better* moral judgments;

(5) even granting *arguendo* the validity of Greene's dual-process model, Greene's normative claims are problematic because his normative theory lacks a deep discussion about the attribution of moral relevance.

More specifically, in Ch. 1 I have drawn a working hypothesis relative to different meanings of 'moral intuition'. In Ch. 2 I have described Greene's descriptive views, reviewed evidence and counter-evidence for his model, and concluded that the latter is the

one that best fits data points stemming from 'trolleyology'. In Ch. 3 I have analyzed several other views in experimental moral psychology that compete with Greene's. I have shown that most of them cannot be ruled out at the moment, so that theories in experimental moral psychology are still underdetermined by the data. The descriptive debate is not going to end any time soon. Furthermore, I have discussed the normative implications these scientists draw from their views, if they do so, and noticed that the meta-normativity problem is a recurring issue for neuromoral theories. In Ch. 4 I have analyzed Greene's neuromoral theory more in detail, claiming that his main argument, the AMIF, does not work even if all of Greene's descriptive claims are granted for the sake of argument. From the descriptive point of view, Greene's view is importantly linked with empirical evidence that stems from the use of hypothetical cases. Since many psychologists and neuroscientists contest the use of hypothetical cases to study moral behavior (§ 2.3.), an important part of Greene's theory stands or falls together with the viability of that methodology. However, Greene's experimental work (and especially Greene et al. 2008, 2009) is praiseworthy and has greatly contributed to gain some initial insight in the workings of the machinery for moral judgments. The normative part of Greene's work, that avowedly follows consequentialist lines, is more problematic than his descriptive work. The AMIF, which is by far the most intriguing and promising argument Greene marshals to attack the deontologist on empirical grounds, is puzzling because Greene has not devoted a specific analysis to judgments about moral relevance, which are the key normative premises of the AMIF itself. The AMIF would work well, if the normative premise could easily be granted. Nonetheless, premises about the relevance of factors are equivalent to moral principles and as likely to be controversial as moral principles. They are not necessarily uncontroversial, as Greene seems to suppose. Moreover, the AMIF runs into further problems concerning the interaction of different factors, the involvement of System1 in the assessment of moral relevance, and the fact that, in a clash between judgments about moral relevance and Haidtian moral intuitions about

cases, it is not clear why judgments about relevance should necessarily prevail as Greene seems to assume.

As to the more general relationship between experimental moral psychology and normative ethics, most neuromoral theories that have been put forth so far have to address the meta-normativity problem, which is a philosophical issue. Another important issue is a philosophical analysis of the attribution of moral relevance to factors. Nonetheless, the relationship will be better understood the more experimentation in moral psychology progresses. At the moment theories are still largely underdetermined by the available data points and consensus must be reached on important methodological quandaries such as hypothetical scenarios. Most arguments trying to derive normative consequences would benefit from both philosophical and empirical work. I have no doubts that experimental moral psychology could prove fruitful to normative ethics, if only to take into account hard-wired tendencies and proclivities of the human mind. It could turn out that improved knowledge of the machinery for moral judgments may help human beings to perform judgments that are better from the prudential point of view, even though this is a thoroughly empirical issue, and besides a difficult one to investigate. It could also turn out that a global EDA may lead to a general anti-realism about evaluative facts, as Street (2006) claims, or it might happen that morality is not a natural kind from the scientific point of view, as Sinnott-Armstrong and Wheatley (2012) suggest. There is still a lot to discover in experimental moral psychology, so that we cannot know how things will turn out to be. However, it is very probable that the set of plausible theories will slowly shrink and that, with less theories on the table, we will be able to understand more clearly what the eventual normative implications might be. It would be helpful to integrate philosophical and experimental research into a common research program, as experimental philosophers are trying to do. Surely armchair theorizing is not very helpful if we want to know whether a moral intuition is widespread or not. At the same time, empirical research can identify psychological processes, such as the Knobe effect, that could undermine the

credibility of some normative ethical claims, such as the DDE (cf. Levy 2011). The future of experimental moral psychology seems to be rosy and I confidently wait for the exciting developments the following decades will undoubtedly yield.

Appendix: Culture – a problem for experimental moral psychology

Experimental moral psychology faces a problem, together with the other 'behavioral sciences' (cognitive science, experimental economics and psychology). It is not a new problem, but it has recently been put back into the limelight by a beautiful paper by Henrich, Heine, and Norenzayan (2010). This problem is a sampling bias. Most experiments in these sciences are carried out on culturally homogeneous samples. As Henrich, Heine, and Norenzayan point out, people in the typical sample for these studies are WEIRD (Western Educated Industrialized Rich Democratic). Arnett (2008) has surveyed the articles of the main peer-reviewed journals in psychology in the 2003-2007 period and has found that 68% of the subjects come from the US. Furthermore, 67% of this US population is composed of university students who take psychology courses. Therefore, the bulk of experimental subjects in the behavioral sciences is composed by a very specific human group: US undergrads in psychology. On the one hand, this is an advantage, because very homogeneous samples allow the attribution of differences in the subjects' behavioral responses to the differences in the experimental conditions (e.g. distinct stimuli), which are manipulated by the researchers. Moreover, university students are easily available, cheap, and permit a fast replication of the experiments. On the other hand, this poses serious questions of generalizability of experimental findings. How can a researcher be sure that the experimental results are valid under different cultural conditions? This risk is particularly serious if we take into account that university students are a very specific sample relative not only to the global human population, but also to the US population. As Rozin (2010) has pointed out, the university student experiences a unique life transition from family life to a peer-centered life. Moreover, they usually earn little or no income, live in a very liberal, educated, and open-minded environment (the

campus), and have not built their own family yet. Therefore, their behavior on several accounts, such as economic decisions, is likely to be different even from that of the average US 30-year old person. Wide variability in economic behavior is also underscored by cross-cultural studies (Henrich et al. 2005) showing that the behavior of university students coming from Western, industrialized countries on some economic games like the UG and the dictator game is very different from the behavior found in many small-scale societies around the globe.

Before dealing directly with cultural variability, it is useful to briefly illustrate the concept of 'culture' I avail myself of. Please take what follows as a working hypothesis. As I take it, 'culture' refers to features of human groups that typically vary according to geographic areas and which depend on social learning. Languages, religions, shared attitudes and beliefs, family structures, and hierarchies are all parts of culture. Culture varies not only moving from one social group to another, but also from an individual to another in the same group[180]. Culture possesses several dimensions, which are notoriously difficult to measure, so that it is much more complex to take this source of variation into account than others, such as for instance a mono-dimensional factor like age. One framework that I find helpful to deal with the complexity of culture is Hofstede's (2001) five dimensional model, which collocates each culture along these dimensions:

1. individualism – collectivism;

2. small – large power distance: It measures the difference in power between the most and the least powerful members of the group. If power distance is large, the leaders of the group are much more powerful than the subordinates. If power distance is small, the leaders of the group are almost on the same level as subordinates;

3. short – long term orientation: to what degree a group considers the remote future when making decisions;

4. weak – strong uncertainty avoidance: how much a group is willing to take up risks;

[180] See for instance Haidt and Graham (2007) and Haidt, Graham, and Joseph (2009) about the different moral foundations used by liberals and conservatives according to Haidt's MFT.

5. masculinity – femininity: here Hofstede uses the Western stereotypes as metaphors, without any commitment about the actual psychology of men and women. Masculinity symbolizes an assertive and competitive stance, whereas femininity indicates a caring and modest attitude. According to this model, every society is characterized by a set of five values that describe its position along the dimensions, but any individual in the society might depart from the group's values. For instance, the United States (US) are considered as one of the most individualistic societies in the world (Henrich et al. 2010), but a single US citizen can endorse collectivist values for a variety of reasons, such as religious tenets or family education.

As to the behavioral sciences more specifically, cross-cultural variation in human psychology is pervasive (Nisbett and Masuda 2003; Norenzayan and Heine 2005) but it is rarely addressed in the behavioral sciences (Henrich, Heine, and Norenzayan 2010; Sears 1986). Cross-cultural variability in psychology corresponds, in some cases at least, to cross-cultural neural variability (for a review about cross-cultural neural variation, see Han and Northoff 2008). Both geographical variability and individual variability have behavioral consequences. For instance, Chua, Boland and Nisbett (2005) have demonstrated that the Americans and the Chinese feature different saccades[181] patterns when they are shown a picture composed by a salient object and a background: the Chinese tend to focus more on the background than the Americans. As to individual variation, priming for individualism or collectivism[182] performed on bicultural individuals, such as Japanese-Americans, modulates both their ways of self-description (general, context-free descriptions vs. contextual descriptions) and the corresponding BOLD signals in areas related to self- representation (Chiao et al. 2009). Lastly, it must be understood that ethnicity is not a synonym of culture, since immigrants retain their ethnicity for some

[181] Saccades are quick and simultaneous movements of both eyes in the same direction. Human beings are usually not aware of performing saccades.

[182] Individualists think that people are independent from each other and that they are characterized by a context-independent set of personality traits. Collectivists see persons as interconnected and describe them as embedded in specific social situations, which constitute a part of their personality.

generations (as long as they mate with other immigrants coming from the same ethnic group), whereas they rapidly lose their original cultural traits (Heine and Lehman 2004). Individual and intra-national variation also prevents us from identifying culture with nationality, even though nationality has a great influence on culture.

Given the intricacies of culture, experimental moral psychologists that are interested in studying the moral judgments of *H sapiens* and not the moral judgments of the US undergrad should be aware of the problem and include cross-cultural experiments into their experimental strategies, in order to check if data are consistent across different cultures. If this is not done, the experiments risk having a low external validity, i.e. they are based on an idiosyncratic sample which is not representative of the general population. In this case the results would tell little about what happens outside the lab.

A further consideration is that cultural variability does not only involve social behaviors like theory of mind and its neural correlates (Kobayashi Frank and Temple 2009) or economic behavior, but has a much broader scope. For instance, at the behavioral level culture influences general strategies of reasoning (Nisbett et al. 2001), the performance on the visual ‘rod-and-frame' task (Kitayama et al. 2003), and the effectiveness of visual illusions (Segall, Campbell, and Herskovits 1968). Since one may understand the aim of the behavioral sciences as describing universal features of human behavior and accounting for those features by means of appropriate theories, experiments that are carried out on a very specific sample are of little utility to the pursuit of such a purpose, at least as long as they are not repeated in different human groups that diverge culturally. It should be noted that universality must not be intended as a binary (0 or 1) variable: there are discrete degrees of universality that can empirically be tested. For instance, a cognitive phenomenon can be present in almost all human groups, but perform different functions in different contexts, or it can be consistently present and robustly perform the same function in all contexts, or it can be present in some groups only. Universality can be tested by means of three kinds of experiments: (1) the two-cultures experiment, (2) the triangulation

study, and (3) the cross-cultural survey (Norenzayan and Heine 2005). In a two-cultures experiment a determined response to an experimental setting is taken into account. Two cultures that differ on many cultural dimensions are examined and the experimenters check whether the effect is conserved. If it is, the experiment provides some evidence for some degree of universality; if it is not, the difference in the behavioral effects of the setting must be traced back to a cultural dimension. But since the two cultures that have been examined differ on many dimensions, identifying the dimension that is responsible for the variation is not straightforward. In order to do so, a triangular study is needed. Such a study must start from a theory that allegedly explains the previously tested effect and that allows researchers to make hypotheses as to which cultural dimension is responsible for the variation. Then the experimenters take into account three cultures that differ from each other along two theoretically relevant cultural dimensions. For instance, if the theory leads to the prediction that dimensions D1 and D2 may be relevant, the cultures will be selected in such a way that cultures C1 and C2 differ on D1, and C1 and C3 differ on D2. If the behavioral difference is spotted between C1 and C2, D1 will be the relevant dimension; if the difference is found between C1 and C3, D2 will be chosen as explanatory instead[183]. Of course, it must be assured that in the different cultures the experimental conditions are interpreted by the subjects in the same way and that the experimental protocol does not change. A cross-cultural survey entails examining many human groups around the world, both in small-scale societies and in urban societies. If no differences are detected, it provides a strong evidence for some degree of universality. Nonetheless, it is costly and difficult to carry out, as experimental rigor cannot be maintained without considerable efforts when different research teams have to work in diverse environments. These cross-cultural investigations can also be carried out by means of meta-analyses, if sufficient data have already been gathered.

[183] A dimension may be one of the constructs identified by Hofstede (2001), such as individualism or power distance.

Furthermore, there are some types of behavioral research in which universality is not an issue, so that idiosyncratic samples can be used without any problems in these cases. As Gächter (2010) correctly points out, US freshmen and sophomores can be very useful to falsify theories in behavioral economics. Falsification is about the research of counterexamples, not about generalizability, so that using undergrads as participants in an experimental study is appropriate when a study aims at falsification. Furthermore, students are cognitively sophisticated as economic theories often require agents to be.

The sampling bias does not involve experimental psychology only, but cognitive neuroscience too. Here, the situation is probably even worse than in experimental psychology. According to Chiao (2009), 90% of the peer-reviewed neuroimaging studies come from Western industrialized countries. But the sampling bias would be a problem only if significant evidence for cultural variability at the neural level has been gathered. Cultural neuroscience provides substantial evidence to this effect. I briefly review part of this evidence (for a more comprehensive review, see Han and Northoff 2008). Gutchess et al. (2006) have used fMRI to identify the neural correlates of a cross-cultural difference between Caucasian Americans and East Asians in image processing: Americans fixate a salient object more than East Asians. This proves that culture modifies neural function when non-verbal stimuli are processed. Zhu and colleagues (2007) have found a differential activation of the Medial PreFrontal Cortex (MPFC), which explains the distinct construal of the self in American and Chinese subjects. In Americans, whose concept of self does not include intimate relatives, the MPFC is activated only in response to judgments concerning the subject herself, whereas among East Asians the same area of the brain also responds to stimuli concerning close relatives, such as the subject's mother. Hedden and colleagues (2008) uncovered the neural correlates of another cross-cultural effect: East Asians are better than Americans at performing tasks that have contextual demands. Conversely, Americans are better than East Asians at ignoring the context if this is required. By means of an fMRI study on Japanese bilinguals, Kobayashi, Glover and

Temple (2006) have found differences in BOLD activation in Japanese and American cultures when subjects perform false belief tasks[184]. Wong and colleagues (2004) have shown in a Positron Emission Tomography (PET) study that the processing of auditory pitch patterns engages the left or right insular cortex when the pitch has a linguistic function, as in Chinese, or not, as in English, respectively. This demonstrates that linguistic variation across cultures correlates with distinct BOLD activation patterns. One can conclude that cross-cultural variation at the neural level concerns both basic brain functions, such as visual processing, and 'higher' functions, such as self-construal. How can this problem be tackled? MRI scanners are expensive and it is difficult to find them in developing countries or to bring them to the homelands of small-scale societies. Conducting cross-cultural experiments in cognitive neuroimaging is therefore difficult. Nevertheless, East Asia provides a rich and industrialized area in which cultural variability relative to the West is still sufficiently high to make two-cultures neuroimaging experiments meaningful. One possible agenda for cultural neuroscience is to look for the neural correlates of the behavioral variation that has been found between East Asia and the US in cultural psychology.

The precise mechanisms by which culture can sculpt the human brain have not been elucidated yet, but the existence of brain plasticity is now an established fact. The brain changes its physical conformation following environmental stimuli or damage. It has been studied both in the context of functional recovery after lesions (Frost et al. 2003; Wall, Xu, and Wang 2002; Winship and Murphy 2009) and in the context of learning (for instance Maguire et al. 2000). Brain plasticity yields a good theoretical framework to create detailed neural explanations of cross-cultural variability in behavior, but cultural neuroscience still has a lot of work to do in order to reach the neurophysiologic level on which small neural populations are taken into account. In addition, there are well-known and warranted ethical limitations to neurophysiologic experimentation in humans.

[184] The false belief task is one of the main tests for theory of mind.

Summing up, experimental psychology and cognitive neuroscience both show a sampling bias: too many experimental participants are drawn from peculiar and idiosyncratic human groups, such as undergrads in psychology. Since both psychological and neural processes feature a remarkable degree of cross-cultural variation, the samples ought to come from diverse cultures in order to have results that can be generalized to *H sapiens* at large. The sampling bias can be overcome, but experimentation will become more costly and complex if researchers want to implement cross-cultural checks. However, these checks seem necessary to make good empirical science and not to trick both funding bodies and the taxpayer. As to experimental moral psychology, some work, such as most of Haidt's, is already cross-cultural (starting from Haidt, Koller, and Dias 1993). Some experimental work is known to survive cross-cultural checks and this holds for most variations of the trolley dilemma (cf. Hauser, Young, and Cushman 2008). Nonetheless, much experimental work and especially neuroimaging experiments need to be checked for cross-cultural variation. If scientists want to elaborate a good theory of *human* (as opposed to *Western* or *WEIRD*) decision making in moral matters, they need to take cultural variation into account. This shows that experimental moral psychology still has a very long future ahead.

Acknowledgements

I am grateful to MVM for having helped me in the darkest of times. This thesis would not have been written without his presence.

Some of the ideas present in Ch. 3 and 4 were developed in conversation with Matteo Mameli, whom I also thank for reading the manuscript and providing helpful feedback.

I thank Paolo Vezzoni and Giovanni Boniolo for kicking me from Continental to Analytic Philosophy.

I thank Bernard Baertschi for reading the whole manuscript and providing insightful feedback.

I thank Patrik Vuilleumier and Corrado Corradi Dell'Acqua for patiently explaining me quite a great deal of stuff about cognitive neuroscience while in Geneva.

I thank Selim Berker and Walter Sinnott-Armstrong for helpful discussion and sharing unpublished material.

I thank Frances Kamm for helpful conversation and offering me a delightful slice of cheesecake near Washington Sq in Manhattan one evening in June 2012.

I finally thank Joshua David Greene for inviting me at Harvard, allowing me to disturb his lab's work for some months, sharing unpublished material (among which a draft of his upcoming book), engaging in thought-provoking conversation, and being extremely kind.

References

Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: core concept or pragmatic understanding? *Analysis*, 64(2), 173–181.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556–74.

Allman, J., & Woodward, J. (2008). What are moral intuitions and why we should care? *Philosophical Issues*, 18, 164–185.

Amit, E., & Greene, J. D. (2012). You See, the Ends Don't Justify the Means: Visual Imagery and Moral Judgment. *Psychological Science*, 23(8), 861–868

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature reviews: Neuroscience*, 7(4), 268–77.

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. *Nature Neuroscience*, 2(11), 1032-7.

Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *The American psychologist*, 63(7), 602–14.

Arvan, M. (2011). Bad News for Conservatives? Moral Judgments and the Dark Triad Personality Traits: A Correlational Study. *Neuroethics*. doi:10.1007/s12152-011-9140-6

— — — (2012). "A Lot More Bad News for Conservatives, and a Little Bit of Bad News for Liberals? Moral Judgments and the Dark Triad Personality Traits: A Follow-up Study". *Neuroethics*. doi:10.1007/s12152-012-9155-7

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences* 17, 1-42

Bartels, D. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, 108, 381–417.

Bartels, D., & Pizarro, D. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, 121(1), 154–161.

Beccaria, C. (1996/1764). *Of crimes and punishments*. J. Grigson (Transl.). New York: Marsilio Publishers.

Bennett, J. (1974). The Conscience of Huckleberry Finn. *Philosophy*, 49(188), 123–134.

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

Bruni, T. (2012). Ventromedial prefrontal cortex lesions and motivational internalism. *AJOB Neuroscience*, 3(3), 19–23.

Campbell, R., & Kumar, V. (2012). Moral Reasoning on the Ground. *Ethics*, 122(2), 273–312.

Caruso, E. M., & Gino, F. (2011). Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior. *Cognition*, *118*(2), 280–5.

Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nature reviews: Neuroscience*, 4, 841–846.

Chaiken, S., & Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York: The Guilford Press.

Chomsky, N. (1965). *Aspects of a Theory of Syntax*. Cambridge (MA): The MIT Press

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35), 12629–33.

Churchland, P. S. (2011). *Braintrust*. Princeton (NJ): Princeton University Press

Ciaramelli, E., Muccioli, M., Làdavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social cognitive and affective neuroscience*, 2(2), 84–92.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–80.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012) Simulating murder: The aversion to harmful action. *Emotion*, 12(1), 2-7

Cushman, F., & Greene, J. D. (2011). Finding faults: How moral dilemmas illuminate cognitive structure. *Social neuroscience*, DOI:10.1080/17470919.2011.614000.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., & Greene, J. D. (2011). Judgment before principle: engagement of the frontoparietal control network in condemning harms of omission. *Social Cognitive and Affective Neuroscience*. Published on line November 22nd, 2011. doi:10.1093/scan/nsr072

Cushman, F., Young, L. & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science*,17(12), 1082-1089.

Damasio, A. (1994). *Descartes' Error. Emotion, reason, and the human brain.* London: Picador.
Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy*, 76(5), 256-282.
— — — (1980a). Reflective equilibrium and Archimedean points. *Canadian Journal of Philosophy*, 10(1), 83-103.
— — — (1980b). On some methods of ethics and linguistics. *Philosophical Studies*, 37(1), 21-36.
— — — (2011). Reflective equilibrium. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, URL = http://plato.stanford.edu/archives/spr2011/entries/reflective-equilibrium
Dawkins, R. (2006/1976). *The Selfish Gene*. 30th Anniversary Edition. New York: Oxford University Press.
Dawson, M. E., Schell, A. M., and Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson (Eds.), *Handbook of Psychophysiology*, 159–181. Cambridge, UK: Cambridge University Press.
Dean, R. (2010). Does Neuroscience Undermine Deontological Theory ? *Neuroethics*, 3, 43–60.
Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The Neuroscientist*, 13(6), 580-93.
De Lazari-Radek, K., & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics*, 123(1), 9–31.
Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: the deliberation-without-attention effect. *Science*, 311(5763), 1005–7.
Dijksterhuis, A., & Nordgren, L. F. (2006). A Theory of Unconscious Thought. *Perspectives on Psychological Science*, 1(2), 95–109.
Doris, J. M. (2002). *Lack of character: personality and moral behavior*. Cambridge (UK): Cambridge University Press.
Dupoux, E., & Jacob, P. (2007). Universal moral grammar: a critical appraisal. *Trends in Cognitive Sciences*, 11(9), 373–8.
Dwyer, S. (2009). Moral dumbfounding and the linguistic analogy: Methodological implications for the study of moral judgment. *Mind & Language*, 24(3), 274–296.
Dwyer, S., & Hauser, M. D. (2008). Dupoux and Jacob's moral instincts: throwing out the baby, the bathwater and the bathtub. *Trends in Cognitive Sciences*, 12(1), 1–2.
Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6(3/4), 169–200.
Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment. *Psychological Science*, 22(3), 295–299.
Feinberg, M., Willer, R., Antonenko, O., & John, O. P. (2012). Liberating Reason From the Passions: Overriding Intuitionist Moral Judgments Through Emotion Reappraisal. *Psychological science*, 23(7), 788–795.
Fellows, L. K. (2006). Deciding how to decide: Ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. *Brain*, 129(4): 944–952.
Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: Evidence from a reversal learning paradigm. *Brain*, 126(8): 1830–1837.
— — — (2007). The role of ventromedial prefrontal cortex in decision making: Judgment under uncertainty or judgment per se? *Cerebral Cortex*, 17(11): 2669–2674.
Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it? *Philosophical Explorations*, 9(1), 83–98.
Fodor, J. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge (MA): MIT Press.
Foot, P. (1972). Morality as a System of Hypothetical Imperatives. *The Philosophical Review*, 81(3), 305-316.
— — — (1978/1967). The Problem of Abortion and the Doctrine of Double Effect. In Foot, P., *Virtues and Vices*, Oxford: Basil Blackwell.
Frost, S. B., Barbay, S., Friel, K. M., Plautz, E. J., & Nudo, R. J. (2003). Reorganization of remote cortical regions after ischemic brain injury: a potential substrate for stroke recovery. *Journal of neurophysiology*, 89(6), 3205–14.
Gächter, S. (2010). (Dis)advantages of students' subjects. What is your research question? *Behavioral and brain sciences*, 33, 92-93.
Gazzaniga, M. (1998). *The Mind's Past.* Berkeley (CA): University of California Press.
— — — (2005). *The Ethical Brain*. New York: The Dana Press.
Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European review of social psychology*, 2(1), 83–115.
— — — (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103(3), 592–596.
— — — (2000). *Adaptive Thinking*. New York: Oxford University Press.
— — — (2005). I think, therefore I err. *Social Research: An International Quarterly*, 72(1), 195–218.
— — — (2008a). Moral Intuition = Fast and Frugal Heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press, 1-26.

— — — (2008b). Reply to Comments. In W. Sinnott-Armstrong (Ed.), *Moral psychology Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press, 41-46.
— — — (2010). Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality. *Topics in Cognitive Science*, 2(3), 528–554.
Gigerenzer, G., & Selten, R. (2001) (Eds.). *Bounded Rationality*. Cambridge (MA): MIT Press.
Gigerenzer, G., Todd, M., & the ABC Research Group (1999). *Simple Heuristics that Make Us Smart*. New York, Oxford University Press.
Glenn A. L., Raine, A., Schug, R. A., Young, L., & Hauser, M. (2009a). The neural correlates of moral decision making in psychopathy. *Molecular Psychiatry,* 14(1), 5–9
— — — (2009b). Increased DLPFC activity during moral decision- making in psychopathy. *Molecular Psychiatry*, 14(10), 910–911.
Glover, J. (1999). *Humanity. A moral history of the twentieth century*. New Haven: Yale University Press
Gohm, C. L. (2003). Mood regulation and emotional intelligence: Individual differences. *Journal of Personality and Social Psychology*. 84, 594–607.
Grafman, J. (1995) Similarities and distinctions among current models of prefrontal cortical functions. *Annals of the New York Academy of Science*, 769, 337–368.
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029-1046
Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2), 366-385.
Greene, J. D. (2003). From neural "is" to moral "ought": what are the moral implications of neuroscientific moral psychology? *Nature Reviews: Neuroscience*, 4(10), 846–9.
— — — (2007). Why are VMPFC more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322-323
— — — (2008a). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 3. The Neuroscience of Morality.* Cambridge (MA): MIT Press, 35-79.
— — — (2008b). Reply to Mikhail and Timmons. In W. Sinnott-Armstrong (Ed.), in *Moral Psychology. Vol. 3. The Neuroscience of Morality*, Cambridge (MA): MIT Press, 105-117
— — — (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
— — — (2010). *Notes on 'The Normative Insignificance of Neuroscience' by Selim Berker.* Published on-line at http://www.wjh.harvard.edu/~jgreene/GreeneWJH/Greene-Notes-on-Berker-Nov10.pdf, accessed August 31st, 2012.
— — — (forthcoming). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. Forthcoming on *Ethics*. Available upon request to the author.
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), 364–71.
Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517–523.
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), 1144–54.
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, 389–400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). A fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105-2108
Gutchess, A. H., Welsh, R. C., Boduroglu, A., & Park, D. C. (2006). Cultural differences in neural function associated with object processing. *Cognitive, affective & behavioral neuroscience*, 6(2), 102–9.
Haidt, J. (2001). The emotional dog and its rational tail. *Psychological Review*, 108, 814-834.
— — — (2003a). The moral emotions. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Eds.), *Handbook of Affective Sciences*. Oxford: Oxford University Press, 852-870.
— — — (2003b). The emotional dog does learn new tricks: A reply to Pizarro and Bloom (2003). *Psychological Review*, 110(1), 197–198.
— — — (2006). *The Happiness Hypothesis: Putting Ancient Wisdom to the Test of Modern Science*. New York: Basic Books.
— — — (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York: Pantheon Books.
Haidt, J. & Bjorklund, F. (2008a). Social Intuitionists Answer Six Questions about Moral Psychology. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press, 181-217.
— — — (2008b). Social Intuitionists Reason, in Conversation. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press, 241-254.
Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, 20(1), 98-116.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and Below Left–Right: Ideological Narratives and Moral Foundations. *Psychological Inquiry*, 20(2-3), 110-119

Haidt, J., & Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55-66.

— — — (2007). The moral mind. How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Lawrence, & S. Stich (Eds.), *The Innate Mind. Vol. 3. Foundations and the Future*. New York: Oxford University Press, 367-391.

— — — (2011). How Moral Foundations Theory Succeeded in Building on Sand : A Response to Suhler and Churchland. *Journal of Cognitive Neuroscience*, 23(9), 2117–2122.

Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.) *Handbook of Social Psychology*, 5th Edition. Hobeken, NJ: Wiley, 797-832.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628.

Han, S., & Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. *Nature Reviews: Neuroscience*, 9(8), 646–54.

Hare, R. M. (1952). *The Language of Morals*. Oxford, UK: Oxford University Press.

— — — (1981). *Moral Thinking*. *Its Levels, Method, and Point.* Oxford, UK: Oxford University Press.

Hauser, M. D. (2006). The liver and the moral organ. *Social Cognitive and Affective Neuroscience*, 1(3), 214–20.

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

Hauser, M. D., Young, L. & Cushman, F. A. (2008). Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2: The Cognitive Science of Morality*. Cambridge (MA): The MIT Press, 107-144.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews: Neuroscience*, 7(7), 523–34.

Hedden, T., Ketay, S., Aron, A., Markus, H. R., & Gabrieli, J. D. E. (2008). Cultural influences on neural substrates of attentional control. *Psychological science*, 19(1), 12–17.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Prehn, K., Schwintowski, H.-P., & Villringer, A. (2005). Influence of bodily harm on neural correlates of semantic and moral decision making. *NeuroImage*, 24(3), 887-97.

Heine, S. J., & Lehman, D. R. (2004). Move the body, change the self: Acculturative effects on the self-concept. In M. Schaller & C. Crandall (Eds.), *Psychological foundations of culture*. Mahwah (NJ): Erlbaum, 305-331.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., et al. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815; discussion 815–55.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, 33, 61–135.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology Section A*, 58(2), 193-233

— — — (2006). Forward inference using functional neuroimaging: dissociations versus associations. *Trends in Cognitive Sciences*, 10(2), 64-69.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks (CA): Sage Publications.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *Science*, 320, 1092–5.

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

Hume, D. (1960/1739). *Treatise on Human Nature.* London: Oxford University Press.

Hursthouse, R. (2002). *On Virtue Ethics*. Oxford (UK): Oxford University Press

Hynes, C. (2008). Morality, Inhibition and Propositional Content. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 3. The Neuroscience of Morality.* Cambridge (Mass): MIT Press, 25-30.

Joyce, R. (2001). *The Myth of Morality*. Cambridge, UK: Cambridge University Press.

— — — (2006). *The Evolution of Morality*. Cambridge (MA): the MIT Press.

Kahane, G. (2011). Evolutionary Debunking Arguments. *Nous*. 45(1), 103–125

Kahane, G., & Shackel, N. (2008). Do abnormal responses show utilitarian bias? *Nature*, 452, E5.

— — — (2010). Methodological Issues in the Neuroscience of Moral Judgment. *Mind & Language*, 25(5), 561–582.

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. (2003). A perspective on judgment and choice. Mapping bounded rationality. *American Psychologist*, 58(9), 697–720.

— — — (2012). *Thinking, Fast and Slow*. London: Penguin Books.
Kahneman, D., & Tversky, A. (1979). An Analysis of Decision Under Risk. *Econometrica*, 47(2), 263–292.
— — — (1996). On the reality of cognitive illusions. *Psychological Review*, 103(3), 582–91; discussion 592–6.
Kamm, F. (1996). *Morality Mortality. Vol. II: Rights, Duties, and Status*. New York: Oxford University Press.
— — — (2007). *Intricate Ethics. Rights, Responsibility, and Permissible Harm*. New York: Oxford University Press
— — — (2009). Neuroscience and Moral Reasoning : A Note on Recent Research. *Philosophy and Public Affairs*, 37(4), 330–345.
Kant, I. (1966/1797). On a supposed right to lie from philanthropy. In: M. J. Gregor (Ed.). *Practical Philosophy*. Cambridge (UK): Cambridge University Press, 611–15.
— — — (1963/1795). Perpetual peace. In L. W. Beck. (Ed.), *On History*. Indianapolis (IN): Bobbs-Merrill, 85-135.
Kass, L. R. (1997). The wisdom of repugnance. *New Republic*, June 2, 17–26.
Kelly, D., Stich, S., Haley, K. J., Eng, S. J., & Fessler, D. M. T. (2007). Harm, Affect, and the Moral / Conventional Distinction. *Mind & Language*, 22(2), 117–131.
Kitayama, S., Duffy, S., Kawamura, T., & Larsen, J. T. (2003). Perceiving an object and its context in different cultures: a cultural look at new look. *Psychological Science*, 14(3), 201–6.
Kitcher, P. (2005). Biology and Ethics. In D. Copp (Ed.). *The Oxford Handbook of Ethical Theory*. Oxford (UK): Oxford University Press, 163-185.
— — — (2006/1993). Four Ways of "Biologizing" Ethics. In E. Sober (Ed.), *Conceptual Issues in Evolutionary Biology*. Cambridge (MA): The MIT Press. 575-586. First published 1993.
Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, 63(3), 190–194.
— — — (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16(2), 309–324.
— — — (2004). Intention, intentional action and moral considerations. *Analysis*, 64(2), 181–187.
— — — (2006). The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies*, 130(2), 203–231.
Kobayashi, C., Glover, G. H., & Temple, E. (2006). Cultural and linguistic influence on neural bases of "Theory of Mind": an fMRI study with Japanese bilinguals. *Brain and language*, 98(2), 210–20.
Kobayashi Frank, C., & Temple, E. (2009). Cultural effects on the neural basis of theory of mind. *Progress in brain research*, 178, 213–23.
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, 908-911.
— — — (2008). Reply to Kahane and Shackel, *Nature*, 452, E5-E6.
Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. (2011). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*. Published on-line July 18th 2011, doi :10.1093/scan/nsr048
Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of Socialization Theory and Research*, Chicago: Rand McNally, 347-480.
Koven, N. S. (2011). Specificity of Meta-emotion Effects on Moral Decision making. *Emotion*, 11(5), 1255–1261.
Kozel, F. A., Johnson, K. A., Grenesko, E. L., Laken, S. J., Kose, S., Lu, X., Pollina, D., Ryan, A., & George, M. S. (2009). Functional MRI detection of deception after committing a mock sabotage crime. *Journal of Forensic Science*, 54(1), 220-231.
Klein, C. (2010). The Dual Track Theory of Moral Decision making: a Critique of the Neuroimaging Evidence. *Neuroethics*, 4(2), 143–162.
Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., et al. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50), 20084-9.
Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27(3), 313–27; discussion 328–76.
Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology*, 25(3), 311–330.
Lassiter, G. D., Lindberg, M. J., González-Vallejo, C., Bellezza, F. S., & Phillips, N. D. (2009). The deliberation-without-attention effect: evidence for an artifactual interpretation. *Psychological Science*, 20(6), 671–5.
Latané, B., & Darley, J. M. (1970). *The Unresponsive Bystander. Why Doesn't He Help?* New York: Appleton-Century-Crofts

Leben, D. (2011). Cognitive Neuroscience and Moral Decision making : Guide or Set Aside? *Neuroethics*, 4(2), 163–174.
Levy, N. (2006a). Cognitive scientific challenges to morality. *Philosophical Psychology*, 19(5), 567–587.
— — — (2006b). The wisdom of the pack. *Philosophical Explorations*, 9(1), 99-103.
— — — (2007). *Neuroethics. Challenges for the 21st Century*. Cambridge (UK): Cambridge University Press
— — — (2011). Neuroethics: A New Way of Doing Ethics. *AJOB Neuroscience*, 2(2), 3–9.
Lillehammer, H. (2010). Methods of ethics and the descent of man: Darwin and Sidgwick on ethics and evolution. *Biology & Philosophy*, 25(3), 361–378.
Lipsey, R. G. (1956). The general theory of the second best. *Review of Economic Studies* 24, 11–32.
Lipton, P. (1991). *Inference to the best explanation*. Oxford: Routledge.
Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature,* 453, 869-878
Machery, E. (forthcoming). *In defense of reverse inference*. Unpublished manuscript.
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.
Mackie, J. (1977). *Ethics: inventing right and wrong*. London: Penguin Books
McIntyre, A. (1982). How moral agents became ghosts. *Synthese*, 53, 295–312.
Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences of the United States of America*, 97(8), 4398–403.
Mameli, M. (2008). On innateness: the clutter hypothesis and the cluster hypothesis. *Journal of Philosophy*, 105(12): 719-736.
Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An Investigation of Moral Judgement in Frontotemporal Dementia. *Cognitive and Behavioral Neurology*, 18(4), 193–197.
Mendez, M. F., Chen, A. K., Shapira, J. S., & Miller, B. L. (2005). Acquired sociopathy and frontotemporal dementia. *Dementia and geriatric cognitive disorders*, 20(2-3), 99–104.
Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, 11(4), 143–52.
— — — (2008). Moral Cognition and Computational Theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 3. The Neuroscience of Morality.* Cambridge (MA): MIT Press, 81-91.
— — — (2011). Emotion, Neuroscience, and Law: A Comment on Darwin and Greene. *Emotion Review*, 3(3), 293–295.
Milgram, S. (1963) Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–78.
Moll, J. & De Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8): 319-321.
Moll, J., De Oliveira-Souza, R., Eslinger, P., Bramati, I., Mourao-Miranda, J., Andreiuolo, P., & Pessoa, L. (2002a). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, 22(7), 2730-2736.
Moll, J., De Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002b). Functional networks in emotional moral and nonmoral social judgments. *NeuroImage*, 16(3), 696-703
Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008a). The Neural Basis of Moral Cognition. Sentiments, Concepts, and Values. *Annals of the New York Academy of Science*, 1124, 161–180
Moll, J., De Oliveira-Souza, R., Zahn, R., & Grafman, J. (2008b). The Cognitive Neuroscience of Moral Emotions. In W. Sinnott-Armstrong (Ed.). *Moral Psychology. Vol. 3. The Neuroscience of Morality*. Cambridge (MA): The MIT Press, 1-17
Moll, J., Paiva, M. L. M. F., Zahn, R. & Grafman, J. (2008c) Response to Casebeer and Hynes. In W. Sinnott-Armstrong (Ed.). *Moral Psychology. Vol. 3. The Neuroscience of Morality*. Cambridge (MA): The MIT Press, 31-33
Moll, J., Zahn, R., De Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The Neural Basis of Human Moral Cognition. *Nature Reviews: Neuroscience*, 6(10), 799-809.
Montague, P., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., … & Fisher, R. E. (2002). Hyperscanning: Simultaneous fMRI during Linked Social Interactions. *NeuroImage*, 16(4), 1159–1164.
Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, 83, 1029–50.
Moretti, L., Dragone, D., & Di Pellegrino, G. (2009). Reward and social valuation deficits following ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, 128–140.
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.
Moore, A. B., Clark, B. A, & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, 19(6), 549–57.
Nadelhoffer, T., & Feltz, A. (2008). The Actor – Observer Bias and Moral Intuitions : Adding Fuel to Sinnott-Armstrong's Fire. *Neuroethics*, 1, 133–144

Narvaez, D. (2008). The Social Intuitionist Model: Some-Counter Intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press, 233-240.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S., […], and Benson, D. F. (1998) Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria. *Neurology*, 51: 1546–1554.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100(3), 530-42.

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19), 11163–70.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, 108(2), 291–310.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. Verbal reports on mental processes. *Psychological Review* 84(3), 231-259.

Nordgren, L. F., Bos, M. W., & Dijksterhuis, A. (2011). The best of both worlds: Integrating conscious and unconscious thought best solves complex decisions. *Journal of Experimental Social Psychology*, 47(2), 509–511.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: what are they and how can we know? *Psychological bulletin*, 131(5), 763–84.

Nozick, R. (1974). *Anarchy, State, Utopia.* New York: Basic Books

— — — (1981). *Philosophical Explanations.* Cambridge (MA): Harvard University Press

Nussbaum, M. (2001). *Upheavals of Thought.* New York: Cambridge University Press

— — — (2004). *Hiding from Humanity*. Princeton (NJ): Princeton University Press

Ogawa, S., Lee, T. M., Kay, a R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24), 9868–72.

O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgment and Decision Making*, 5(7), 547–554.

Paolacci, G., Chandler, J. & Ipeirotis, G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5), 411-419.

Parfit, D. (1978). Innumerate ethics. *Philosophy and Public Affairs*, 7(4), 285–301.

Parkinson, C., Sinnott-Armstrong, W., Koralus, E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience*, 23(10), 3162–3180.

Paxton, J. M., & Greene, J. D. (2010). Moral Reasoning: Hints and Allegations. *Topics in Cognitive Science*, 2(3), 511–527.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1), 163–77.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews: Neuroscience*, *9*(2), 148–58.

Petrinovich, L., & O'Neill, P. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology*, 17, 145–171.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An Empirical Study of Moral Intuitions : Toward an Evolutionary Ethics. *Journal of Personality and Social Psychology*, 64(3), 467–478.

Pinillos, N., Smith, N., & Nair, G. (2011). Philosophy's new challenge: experiments and intentional action. *Mind & Language*, 26(1), 115–139

Pinker, S. (1997). *How the Mind Works*. New York: Norton.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*, 110(1), 193–196.

Pizarro, D. A. & Uhlman, E. L. (2005). Do normative standards advance our understanding of moral judgment? *Behavioral and Brain Sciences*, 28(4): 558-559

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

— — — (2010). Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? *Perspectives on Psychological Science*, 5(6), 753–761.

Poldrack, R. A., & Wagner, A. D. (2004) What can neuroimaging tell us about the mind? *Current Directions in Psychological Science*, 13(5), 177-181

Price, C. J., & Friston, K.. J. (2002). Degeneracy and cognitive anatomy. *Trends in Cognitive Sciences*, 6(10), 416–421.

— — — (2005). Functional ontologies for cognition : The systematic definition of structure and function. *Cognitive Neuropsychology*, 22(3-4), 262-275

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

— — — (2008). Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Mikhail. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press, 157-170.

Pust, J. (2012). Intuition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2012 Edition)*. Accessible at http://plato.stanford.edu/archives/win2012/entries/intuition/. Accessed February 6th, 2013.

Quinn, W. (1989a). Actions, intentions, and consequences: The doctrine of doing and allowing. *The Philosophical Review*, 98(3), 287–312.

— — — (1989b). Actions, intentions, and consequences: the doctrine of double effect. *Philosophy and Public Affairs*, 18(4), 334–51.

Rawls, J. (1951). Outline of a decision procedure for ethics. *The Philosophical Review*, 60(2), 177-197.

— — — (1999/1971). *A Theory of Justice.* Revised edition. Cambridge (MA): Harvard University Press. First published 1971.

Roskies, A., & Sinnott-Armstrong, W. (2008). Between a Rock and a Hard Place: Thinking about Morality. *Scientific American*, July 29th 2008 – Available at http://www.scientificamerican.com/article.cfm?id=thinking-about-morality (accessed Sep, 3rd 2012)

Ross, W. D. (1930/2002). *The Right and the Good.* Oxford (UK): Oxford University Press. First published 1930.

Rozin, P. (2010). The weirdest people in the world are a harbinger of the future of the world. *Behavioral and Brain Sciences*, 33, 108-109

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76(4), 574-586.

Ruse, M. (1986). *Taking Darwin seriously.* Amherst (NY): Prometheus Books.

Ruse, M. & Wilson, E. O. (2006/1986). Moral philosophy as applied science. In E. Sober (Ed), *Conceptual Issues in Evolutionary Biology*. Cambridge (MA): The MIT Press. 555-573 First published 1986.

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review*, 110(1), 145–172.

Saltzstein, H. D., & Kasachkoff, T. (2004). Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique. *Review of General Psychology*, *8*(4), 273–282.

Saver, J. L., & Damasio, A. R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, 29(12), 1241–1249.

Savulescu, J. & Sandberg, A. (2008). Neuroenhancement of love and marriage: the chemicals between us. *Neuroethics* 1: 31-44

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4), 1835-1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43, 1391–1399..

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

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an FMRI investigation. *Journal of Cognitive Neuroscience*, 18(5), 803–17.

Schaich Borg, J., Lieberman, D., & Kiehl, K. (2008). Infection, incest, and iniquity: Investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, 20(9), 1529–1546.

Scherer, K. R. (2009a). Emotions are emergent processes : they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society of London: Series B*, 364, 3459–3474.

— — — (2009b). The dynamic architecture of emotion : Evidence for the component process model. *Cognition and Emotion*, 23(7), 1307–1351.

Schnall, S., Benton, J., & Harvey, S.(2008). With a Clean Conscience. Cleanliness Reduces the Severity of Moral Judgments. *Psychological Science*, 19, 1219-1222.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as Embodied Moral Judgment. *Personality and Social Psychology Bulletin*, 34, 1096–1109.

Sears, D. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51(3), 515–530.

Segall, M., Campbell, D. T., & Herskovit, M. J. (1968). The Influence of Culture on Visual Perception. *Studies in Art Education*, 10(1), 68.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, 132(3): 617–627.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: The role of the right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience*, 15(3): 324–337.

Shweder, R. A., Much, N. C., Mahapatra, M., and Park, L. (1997). The "Big Three" of Morality (Autonomy, Community, Divinity) and the "Big three" explanations of suffering. In A. Brandt & P. Rozin, (Eds.), *Morality and Health*. New York: Routledge, 119-169.

Sidgwick, H. (1907). *The Methods of Ethics*. 7th Edition. Hackett: Indianapolis.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, 99-118

— — — (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129-138
Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, 1(3), 229–243.
— — — (1974). Sidgwick and reflective equilibrium. *The Monist*, 58, 490-517.
— — — (1981). *The Expanding Circle*. Princeton (NJ): Princeton University Press.
— — — (2005). Ethics and intuitions. *The Journal of Ethics*, 9, 331–352.
Sinnott-Armstrong, W. (2006). *Moral Skepticisms*. Oxford, UK: Oxford University Press.
— — — (2008a). Framing Moral Intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*, Cambridge (MA): The MIT Press, 47-76.
— — — (2008b). Is moral phenomenology unified? *Phenomenology and Cognitive Science*, 7, 85-97.
— — — (2008c). How to apply generalities: Reply to Tolhurst and Shafer-Landau. In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*, Cambridge (MA): The MIT Press, 97-105.
Sinnott-Armstrong, W., & Wheatley, T. (2012). The disunity of morality and why it matters to philosophy. *The Monist* 95(3), 355-377.
Sober, E., & Wilson, D. S. (1999). *Unto Others*. Cambridge (MA): Harvard University Press.
Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach*. Oxford: Basil Blackwell.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5), 645–65; discussion 665–726.
Sterelny, K. (2010). Moral Nativism: A Skeptical Response. *Mind & Language*, 25(3), 279–297.
Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, 127, 109–166.
Strick, M., Dijksterhuis, A., & van Baaren, R. B. (2010). Unconscious-thought effects take place off-line, not on-line. *Psychological Science*, 21(4), 484–8.
Suhler, C. L., & Churchland, (2011). Can innate, modular "foundations" explain morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience*, 23(9), 2103–16.
Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, 28, 531–573.
— — — (2008). Fast, frugal and (sometimes) wrong. In W. Sinnott-Armstrong, (Ed.), *Moral Psychology. Vol. 2. The Cognitive Science of Morality*. Cambridge (MA): MIT Press. 27–31.
Sunstein, C. R. & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review* 70(4), 1159-1202.
Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, 119(3), 454–458.
Taurek, J. M. (1977). Should the numbers count? *Philosophy and Public Affairs*, 6(4), 293–316.
Tersman, F. (2008). The reliability of moral intuitions: A challenge from neuroscience. *Australasian Journal of Philosophy*, 86(3), 389–405.
Theysohn, J. M., Maderwald, S., Kraff, O., Moenninghoff, C., Ladd, M. E., & Ladd, S. C. (2008). Subjective acceptance of 7 Tesla MRI for human imaging. *Magnetic Resonance Materials in Physics, Biology and Medicine* (MAGMA), 21(1-2), 63–72.
Thomas, B. C., Croft, K. E., & Tranel, D. (2011). Harming kin to save strangers: further evidence for abnormally utilitarian moral judgments after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, 23(9), 2186-96.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59: 204-217
— — — (1985). The Trolley Problem. *Yale Law Journal*, 94: 1395-1415
— — — (2008). Turning the Trolley. *Philosophy and Public Affairs*, 36(4): 359-374
Todd, M., & Gigerenzer, G. (2007). Mechanisms of ecological rationality: heuristics and environments that make us smart. In R. Dunbar & L. Barrett (Eds.), *Oxford Handbook of Evolutionary Psychology*. Oxford, UK: Oxford University Press.
Trivers, R. (1971). The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*, 46(1), 35–57.
Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. New York: Cambridge University Press.
Tversky, A, & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–31.
— — — (1981). The Framing Of Decisions and the Psychology of Choice. *Science*, 211(4481), 453–458.
— — — (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4), 293–315.
Unger, P. (1996). *Living High and Letting Die. Our Illusion of Innocence.* New York: Oxford University Press.
Valdesolo, P., & Desteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, 17(6), 476–477.
Van Thiel, G. J. M. W, & Van Delden, J. J. M. (2010). Reflective Equilibrium as a Normative Empirical Model. *Ethical Perspectives*, 2, 183-202.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton (NJ): Princeton University Press.
Wall, J., Xu, J., & Wang, X. (2002). Human brain plasticity: an emerging view of the multiple substrates and mechanisms that cause cortical changes and related sensory dysfunctions after injuries of sensory inputs. *Brain Research Reviews*, 39, 181–215.

Wheatley, T., & Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, 16(10), 780–784.
Williams, G. C. (1966). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton (NJ): Princeton University Press.
Wilson, E. O. (1975). *Sociobiology. The New Synthesis*. Cambridge (MA): Harvard University Press.
— — — (1979). *On Human Nature.* Cambridge (MA): Harvard University Press.
Winship, I. R., & Murphy, T. H. (2009). Remapping the somatosensory cortex after stroke: insight from imaging the synapse to network. *The Neuroscientist*, 15(5), 507–24.
Wong, C. M., Parsons, L. M., Martinez, M., & Diehl, R. L. (2004). The role of the insular cortex in pitch pattern perception: the effect of linguistic contexts. *The Journal of Neuroscience*, 24(41), 9153–60.
Woodward, J., & Allman, J. (2008). Moral intuition: its neural substrates and normative significance. *Journal of Physiology, Paris*, 101(4-6), 179–202.
Yarkoni, T., Poldrack, R. A., Essen, D. C. V., & Wager, T. D. (2010). Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences*, 14(11), 489–496.
Yarkoni, T., Poldrack, R. A., Nichols, T. E., Essen, D. C. V., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8), 665-670.
Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(15), 6753-6758.
Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, 104(20), 8235–40.
Young, L., & Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social neuroscience*, 7(1), 1–10.
Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, 40(4), 1912–20.
— — — (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, 21(7), 1396–405.
Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15), 6430-6435.
Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *NeuroImage*, 34(3), 1310–6.
Zimbardo, P. G. (2007). *The Lucifer Effect*. New York: Random House.