EVALUATIVE FOCUS:

A DUAL-PROCESS VIEW OF MORAL JUDGMENT

Ivar Allan Rodriguez Hannikainen

A thesis submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Philosophy

University of Sheffield

February 2014

Abstract

In this dissertation, I aim to develop an empirical account of moral judgment. Chapter 1 lays some philosophical and methodological groundwork. Next, in Chapters 2 and 3, I review and critically discuss past literature on moral judgment. In recent decades, automaticity research has led to the view that our social judgments are conducted automatically, and uncontrolled by conscious reasoning. In Chapter 2, I push back against this view, arguing that moral judgments are readily shaped by reasoning processes. Next, in Chapter 3, I differentiate a few empirical claims about the relationship between affective processes and moral judgment, and I arbitrate between them. I then aim to characterize the psychological processes that cause these affective responses, arguing for the involvement of a sensory and motor simulation of the behavior. This exercise gives rise to new empirical hypotheses, which are then tested in Chapters 4 and 5. In Chapter 4, I present a collaborative, empirical study that examines the aversions to harmful and disgusting behavior. Our results suggest that – across both purity and harm –condemnation of immoral behavior arises principally from a personal aversion to performing the target action, which shapes third-party judgments through the partly unconscious simulation of the agent's perspective. In Chapter 5, I present some analyses that examine the broader influence of evaluative focus on moral and political attitudes. Finally, Chapter 6 argues that the proposed psychological account of moral judgment is consistent with evidence from a wider range of disciplines, from neuroscience and animal cognition, to evolutionary theory and sociology of religion.

Word count: 55,397

Acknowledgements

First and foremost, I am grateful to the University of Sheffield for their generous financial support throughout my doctoral studies. I am equally grateful to Stephen Laurence and Fiery Cushman for their invaluable mentorship and support. I would also like to especially mention Ryan Miller for a fruitful collaboration, and James Lenman for his thoughtful comments on previous drafts.

I have been lucky to take part in three exceptional research environments: the Philosophy Department at Sheffield, the Moral Cognition Lab at Harvard, and the Moral Psychology Research Lab at Brown. Each of these has afforded me countless valuable discussions. I will name a few whose feedback has been particularly instructive: Bernardo Aguilera, Dan Bartels, Alek Chakroff, Ryan Doran, James Dungan, Josh Greene, Jon Haidt, Josh Rottman, and Liane Young.

Finally, I should also thank Regan Bernhard, Kyle Dillon, and Bay McCulloch for their assistance in conducting the experimental research reported in Chapters 4 and 5, and my family and friends whose relentless encouragement and support I could not have done without.

Table of Contents

Chapter 1. Introduction 1

1.1. Descriptive philosophy and naturalism 1

1.2. The psychology of moral judgment 5

1.3. Naturalizing moral judgment 10

1.4. Conclusion 13

Chapter 2. Dual process theories and the role of reasoning in moral judgment 17

2.1. Introduction 17

2.2. Haidt’s social intuitionist model 20

2.3. Greene’s dual process theory 22

2.4. The role of controlled processing 23

2.5. Is welfare calculus affectively motivated? 39

2.6. Conclusion 47

Chapter 3. Mental simulation and the spontaneous evaluation of moral events 48

3.1. Introduction 48

3.2. Characterizing the relationship of affect to moral judgment 49

3.3. Further back in the causal chain 60

3.4. Simulation hypothesis: an alternative account? 74

3.5. Conclusion 79

Chapter 4. Testing the hypothesis of simulated aversion 81

4.1. Introduction 81

4.2. Are agent and victim simulation tractable in moral judgment? 85

4.3. Experiment 1: Does the condemnation of purity rely on simulated aversion? 93

4.4. Experiment 2: Does the condemnation of harm rely on simulated aversion? 98

4.5. Discussion 103

4.6. Perspective induction via narration 106

4.7. Experiment 3: Does perspective-taking influence condemnation of purity violations? 107

4.8. Experiment 4: Does perspective-taking influence condemnation of harm? 110

4.9. Discussion 115

4.10. General Discussion 116

Chapter 5. The influence of evaluative focus on moral and political attitudes 120

5.1. Introduction 120

5.2. Analysis 1: Focus- versus content-based accounts of moral disagreement. 129

5.3. Analysis 2: Demographic predictors of evaluative focus. 133

5.4. Analysis 3: The influence of cognitive style on moral evaluation. 136

5.5. Analysis 4: Evaluative foci predict patterns of moral foundations. 141

5.6. Analysis 5: Building a model of politics, moral foundations and evaluative foci. 146

5.7. General discussion 149

Chapter 6. The theory of evaluative focus in perspective. 154

6.1. Introduction 154

6.2. Neuropsychology 158

6.3. Developmental acquisition of aversions to actions 160

6.4. Innate predisposition: the case of incest 164

6.5. Aversions in practice: Behavioral choice and reinforcement learning models 167

6.6. Evaluative simulation in action 173

6.7. Conclusion 184

References 187

Appendices 208

Appendix A 208

Appendix B 209

Appendix C 213

Appendix D 220

Appendix E 232

Supplementary Analyses 233

Chapter 1. Introduction

1.1. Descriptive philosophy and naturalism

Moral philosophers have typically sought to answer questions about how it is we *should* live ethically-speaking: What are our duties and responsibilities towards others? How might we determine what the right thing to do is? These efforts have yielded a variety of *normative* ethical theories, such as virtue ethics (Aristotle, 1991), deontology (Kant, 1785/1964), and utilitarianism (Mill, 1863), each of which provides its own guidelines for differentiating morally right conduct from morally wrong conduct.

Alongside this normative ethical aim, a parallel enterprise has focused on understanding the status of morality more broadly as an object of study. Is morality a kind of science, and so a matter of discovering objective facts about the moral truth? If so, what kind of facts are moral facts, and how do we ordinarily acquire knowledge about them? Did morality evolve as a natural phenomenon, or is it a cultural invention? What does it mean to say that something is “morally right/wrong”? Since Plato, these sorts of *descriptive* questions have occupied philosophers who aimed to understand the place of moral phenomena within the natural and social world.

Hume famously pronounced these to be two fundamentally distinct projects in moral philosophy, and warned us that (descriptive) “observations concerning human affair” could not by themselves entail any claims about the (normative) moral truth:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when all of a sudden I am surprised to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it (Hume, 1739, 335).

Ever since, many moral philosophers have echoed the idea that knowledge of all the relevant descriptive facts cannot by itself entail any normative conclusions. We may know all the facts about the natural laws that ensure that a particular course of action, e.g., swinging a hammer at an old lady, gives rise to some needless pain. And yet, to ground the normative claim that "one ought not to swing the hammer", we need an additional assumption, along the lines of (N1) *one ought not cause needless pain to others*. We may try to derive this latter claim from natural facts, but this is difficult to do also. For instance, we may try to derive N1 from certain descriptive facts like (F1) *causing needless pain tends to decrease the welfare of society*, or (F2) *rational people typically reject norms that allow inflicting needless pain*. However, this step merely creates new normative assumptions in each case, i.e., that (N1') *one ought not decrease the welfare of society* and (N2') *one ought to do whatever it is rational to accept in one's system of norms*. So ultimately, with some exceptions (see Searle, 1964), philosophers have been convinced that normative standards like (N1') and (N2') cannot derive from a descriptive understanding of the relevant facts alone. This insight, credited to Hume, is summarized by the mantra "you cannot derive an ought from an is", and has established an impervious divide between normative and descriptive projects in ethics. My dissertation will fall squarely on the descriptive side, and I will heed Hume's advice to make no claims about what we *ought to do* on the basis of this work.

1.1.1. Naturalist metaphysics

How to proceed with a descriptive project depends critically on one's metaphysical views. In the following paragraphs I will put forth my own metaphysical assumptions, which in philosophical circles are known as *naturalistic* (Papineau, 1993; Sellars, 1956). It is useful to distinguish the *ontological* from the *methodological* dimensions of naturalism, though I will be adopting both. First of all, I will make the following ontological assumption*:*

(1) Ontological naturalism: Everything that exists is constituted of (/reducible to) physical, spatiotemporal objects and/or properties.

This view rules out the existence of immaterial phenomena, i.e., any supernatural or non-natural objects (or properties) that are not constituted by physical, spatiotemporal objects (or properties). For instance, in making this assumption, I am granting that people's moral judgments must be reducible to the psychological and neural states that instantiate them. Similarly, the moral properties of an event (if they are to exist at all) must ultimately be reducible to its physical, spatiotemporal properties.

One reason to adopt a naturalist ontology has to do with advocating the causal efficacy of moral phenomena. A popular view has it that the physical world is *causally closed*, meaning that every physical effect has a sufficient physical cause (Kim, 1993). So if we want to retain the assumption that an agent's moral judgments, can at least sometimes cause changes in her behavior (e.g., if a moral concern for the environment leads one to vote for a candidate with a better environmental campaign), then we shall have to accept that moral judgments are physically constituted. Therefore a naturalist ontology of this sort rules out any metaethical view which posits moral phenomena that are not physically constituted (e.g., Adams, 2002).

A second, closely-related assumption that I will be making concerns the methodology by which we discover descriptive facts and construct an ontology:

(2) Methodological naturalism: Whatever exists can be observed using the best standard observational procedures in the natural sciences.

This assumption requires that the ontology of our philosophical theories derive from the entities, relations and properties observed using scientific methods. That is, whatever exists in philosophical theories must be documented by some natural and/or social science.

So according to methodological naturalism, descriptive philosophy is in many ways continuous with science. This has engendered some worries about the distinctiveness of philosophy, which I will attempt briefly to assuage. In short, even while both rely on empirical data, philosophy and the sciences answer different kinds of questions using these data. For example, while the primatologist may have an inherent interest in the bonobo's social behaviors, the philosopher seeks to understand the same body of work in order to answer *broader* and more *resilient* questions about, say, the nature of mind-reading. Because of these characteristics of philosophical issues, the philosopher will have to critically examine evidence from across numerous disciplines – primatology, but also linguistics, psychology, neuroscience, cultural anthropology perhaps – and synthesize their relevant findings so as to answer or at least constrain the space of plausible theories about the philosophical question at stake. This is analogous to the way in which scientific theories must integrate and synthesize evidence across numerous experiments, in order to construct scientific theories (see Figure 1).

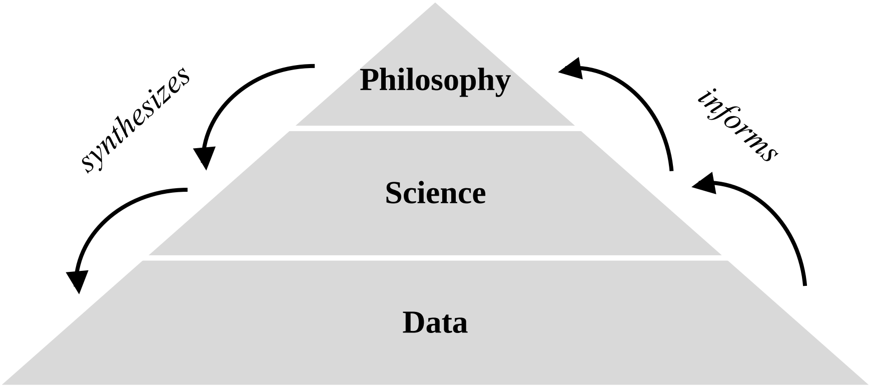


Figure 1. Methodological naturalism.

So, for naturalists, philosophical theories are still synthetic theories about the natural world, informed by research across multiple scientific disciplines and ultimately answerable with empirical data. This inherent dependency on science has caused discomfort among philosophers who perceive the spread of naturalism as a threat to their discipline. The worry is that, on the naturalist picture, there are no skills or knowledge that are unique to philosophy. Scholars in other disciplines have all the requisite tools to do philosophy, and if anything it is philosophers who lack the technical skills required to collect new data with which to inform their theories. However, I have shown that, in adopting methodological naturalism, although we eliminate any fundamental methodological differences between science and philosophy, there are differences in their *theoretical objectives* which suffice to make philosophy a unique intellectual endeavor.

1.2. The psychology of moral judgment

In my dissertation, I aim to adopt this naturalistic approach to the examination of issues in descriptive ethics. To provide a *comprehensive* theory of morality would require a cohesive explanation of diverse aspects: how the mental practice of making moral judgments (*moral psychology*), links up with the meaning of our moral language (*moral semantics*), while explaining how it is that we learn or acquire these moral views (*moral epistemology*). This much I cannot take on within the span of this dissertation, so I will delve into one corner of this puzzle: the psychology.

A central ambition of philosophical moral psychology is to give an account of what our minds are up to when we form and issue moral judgments. First, I would like to stop to consider what is meant by ‘moral judgment’. Moral judgment, as I will be understanding it throughout my dissertation, is (i) a *mental state*. In other words, it is a psychological state that an individual can instantiate at a given time: e.g., feeling tired, believing there is coffee in the mug, or thinking about a pink elephant. In this sense, moral judgments are not statements, or any kind of public, linguistic expressions of a moral judgment (although these are, of course, related). They are (ii) *evaluative*; i.e., they ascribe value – rightness/wrongness, goodness/badness, permissibility, obligatory-ness, forbidden-ness, or whatever. These value ascriptions are directed at (iii) people's *behavior*. They ascribe moral value to things that people might *do*, like donating part of one's income to charity or urinating in public. Related judgments about the moral character of others, or about whether agents ought to be held accountable for their actions, or about deserved punishment for a given offense fall outside the range of this study.

Whose moral judgments are relevant to this study? The moral judgments of anyone with a basic aptitude with moral concepts are considered informative. As such, the theoretical conclusions that I defend throughout are based on inductive claims. They describe probabilistic relations that have been observed by sampling from this large, target class of folk moral judgments. This leaves ample room for individual variability in people's moral judgment, and renders a claim about moral judgment compatible with the existence of numerous counterexamples.

Having cleared this up, let us take a look at the historical perspectives on the debate concerning the psychology of moral judgment. Two fundamental camps derive from a popular distinction, attributed to Hume (1739), between *cognitive* and *non-cognitive* mental states (see also Smith, 1987). On one hand, cognitive mental states attempt to represent the world as it is (e.g., believing one will win the lottery this week). Therefore, cognitive mental states should be responsive to evidence that bears on the truth or falsity of their propositional content. That is, the apprehension of a fact that is contrary to one's belief ought to count as evidence against the belief. On the other hand, we have non-cognitive mental states, like emotions, desires, hopes, fears, etc (e.g., hoping one will win the lottery this week). Rather than representing the world, non-cognitive mental states express the agent's emotions or attitudes towards it. Because non-cognitive states do not *represent* the world, facts that are contrary to their propositional content do not count as evidence against them.

The cognitive/non-cognitive distinction has proven useful in explaining people's intentional behavior. Essentially, cognitive mental states (e.g., the belief that the corner store has ice cream) can be paired with non-cognitive mental states (e.g., the desire for ice cream) in order to generate plans and intentional actions (e.g., going to buy ice cream). In addition, as I said, we can mount the primary camps in philosophical moral psychology onto this distinction: cognitivists about moral judgment argue that moral judgments are primarily belief-like, while non-cognitivists argue they are composed principally of some emotion or other non-cognitive attitude. We shall look closely at these opposing philosophical perspectives in the following sections.

1.2.1. Cognitivism

For cognitivists, moral judgments are beliefs that represent the actions or events in question as instantiating certain moral properties (i.e., wrongness, obligatory-ness, or whatever). These moral properties are then determined by, for example, what rational agents would accept as their duty, or what actions bring about the best consequences for aggregate welfare, and so on. Immanuel Kant (1785/1964) famously defended cognitivism; and more recently, Christine Korsgaard (1996) and Michael Smith (1994) have put forth versions of this view. Certain desiderata, however, are shared by cognitivist proposals across the board. For cognitivists, moral judgments are essentially an *exercise of reason*, in two concrete ways:

First, because moral judgments represent actions and/or events, they are *truth evaluable*. That is, moral judgments are true (/false) if the actions or events in question exhibit (/do not exhibit) the properties ascribed to them. This desideratum is, of course, compatible with all moral beliefs being false (see *error theory* in Mackie, 1977).

Second, for rationalist cognitivists, moral judgments are also *inferential*. That is, they are derived from general moral principles along with facts about the specific action or event. For example:

1. General moral principle: Intentionally killing a sentient being is *prima facie* morally wrong.

2. Context-specific fact: Eating meat requires intentionally killing a sentient being.

c. Context-specific moral judgment: Eating meat is *prima facie* morally wrong.

Cognitivism has some intuitive plausibility. One reason to believe that moral judgments are cognitive has to do with the pervasiveness of moral disagreement. I.e., it is obvious that people disagree vigorously about moral issues, and it would seem that disagreement implies truth-evaluability.

Consider a prototypical non-cognitive mental state, such as desiring candy, being afraid of spiders, or wishing to be a famous rock star. These mental states are not truth evaluable in the relevant sense. An audience may question the propriety, normalcy of the agent's attitude, or simply not share the attitude in question, but none of this is inconsistent with the agent's having that attitude. In other words, non-cognitive attitudes are not matters that we can disagree about in the full sense. By contrast, cognitive mental states *are* typically truth evaluable. If an agent has a belief about some particular matter and someone else has a different (i.e., inconsistent) belief about it, it is appropriate and common for them to disagree with each other. This may involve bringing evidence or reasons to bear to justify one's belief, and perhaps also to convince one another of the truth of one's belief. So, in sum, the fact that people disagree and argue profusely about matters of right and wrong (consider the prolific debates surrounding euthanasia, abortion and so on) suggests that moral judgments may be truth evaluable, and therefore, cognitive.

1.2.2. Non-cognitivism

Meanwhile, *non-cognitivist* philosophers since Hume argued that, rather than representing actions as having certain properties, moral judgments primarily involve the expression of an emotion or attitude in relation to the action (Hume, 1751/1894; Ayer, 1952; Hare, 1952). Within non-cognitivism, there is a marked diversity concerning precisely which attitudes are conveyed in moral judgment: sentiments of moral approval and disapproval, the prescription of the behavior in question (Stevenson, 1944), or one's personal acceptance of a norm governing the action (Gibbard, 1990) to list a few examples. And yet, in different ways, most non-cognitivists accept Hume's classic claim that moral attitudes "cannot be the work of the judgement, but of the heart; and is not a speculative proposition or affirmation, but an active feeling or sentiment" (Hume, 1751/1894, 290).

Ayer then furthered the case for non-cognitivism through an analysis of moral language. He pointed out that moral statements like "You acted wrongly in stealing that money" do not add any propositional content to the statement "You stole that money". Rather, they add the speaker's attitude of condemnation towards the action:

> It is as if I had said, "You stole that money" in a peculiar tone of horror, or written it with the addition of some special exclamation marks. […] If now I generalise my previous statement and say, "Stealing money is wrong", I produce a sentence that has no factual meaning -- that is, expresses no proposition that can be either true or false. […] I am merely expressing certain moral sentiments (Ayer, 1952, 107).

Some intuitive considerations favor non-cognitivism also. In particular, numerous philosophers have noted the fact that when people make moral judgments they tend to feel at least partly (and defeasibly) motivated to act in accordance with their judgment. This intuition is not universally shared (see Copp, 1997), but a great many of us find it difficult to conceive of ordinary people making sincere moral judgments while failing to have any disposition or motivation whatsoever to act accordingly. This feature of moral practice (known as *motivational internalism*) has led numerous philosophers toward non-cognitivist views of moral judgment. This is because it is widely accepted that (outside the moral domain) beliefs do not *by themselves* motivate action. My believing that the nearest gas station sells ice cream doesn't by itself motivate me to do anything about this. By contrast, my desire for ice cream, even in the absence of any associated beliefs (where it is available, how much it is healthy to eat, etc.), is enough to motivate me to act, perhaps by leading me to look for some ice cream. In other words, non-cognitive mental states like desires, hopes and emotions do have motivational force. Therefore, if we share the intuition that moral judgments are typically motivating, we might be inclined *prima facie* to favor a non-cognitivist account of moral judgment.

1.3. Naturalizing moral judgment

As we saw, both cognitivism and non-cognitivism each have some intuitive plausibility. At the same time, I highlighted certain aspects of moral practice that they cannot explain so easily. Throughout the 20th century, philosophers concerned with developing a descriptive theory of morality were involved in an ongoing exchange of intuition pumps and intricate arguments intended to favor their particular versions of one of these broad-stroked accounts: Non-cognitivists have argued that disagreement is possible even in a non-cognitive account of moral judgment, cognitivists have argued that moral judgments can be motivating in a cognitivist account of moral judgment, and so on. New theoretical perspectives have emerged from this exchange, many of which blur the cognitive/non-cognitive distinction I outlined above.

However, from the naturalistic perspective I am adopting, these intuitive considerations cannot count as evidence for a descriptive theory. Persuasive arguments and intuitions undoubtedly *motivate* the hypotheses we form, but they cannot provide evidence for them. As I said, for the naturalist, philosophical theories are synthetic theories about the natural world. So the truth about moral judgment (i.e., which version of cognitivism, or non-cognitivism is true) depends not on the success of philosophical arguments but rather exclusively on empirical facts about the phenomenon of moral judgment.

The experimental approach to moral psychology has engendered a third family of accounts, as an alternative to cognitivism and non-cognitivism. These accounts are referred to as *intuitionism* in moral psychology circles (not to be equated with intuitionism as understood in philosophy circles), in that they posit non-affective intuitions – which can be classified neither as cognitive, nor as non-cognitive – as the basis of moral judgment (Huebner, Dwyer, & Hauser, 2009; Mikhail, 2000, 2007).

Intuitions differ in some respects from non-cognitive mental states, like emotions. For instance, the content of an intuitive mental state is a proposition, in our case, the ascription of a moral property to an action or event (see Huemer, 2005). So, unlike non-cognitive mental states, intuitions are representational. Yet, at the same time, intuitions are also different from cognitive mental states. Whereas beliefs are normally inferred from reasons, other beliefs, observations and so on, intuitions are typically *non-inferential* and *self-evident* like sense perceptions (Kornblith, 1998; Gopnik & Schwitzgebel, 1998). According to intuitionists when we observe someone killing an innocent person, for example, we have the spontaneous and self-evident intuition that this is morally wrong, without inference and by mere observation. So, in essence, for intuitionists, moral judgment is the immediate perception of the moral value ascribed to an action or event.

1.3.1. Characterizing intuition and affect

Whereas for intuitionists moral judgment primarily *describes* or *ascribes* moral properties to the action under evaluation, for non-cognitivists it primarily expresses the evaluator's attitude of disapproval or approval related to said action. This difference is worthy of more detailed attention, so in what follows I will clarify two fundamental differences between the concepts of moral intuition and moral affect. First:

Intuitions necessarily have propositional content, but affect does not.

Intuitions are the kind of mental state that carries propositional content (i.e., the intuition that *p*). These propositions need not be consciously held to be true; instead they may dispose one to believe *that p*, or to enact behavioral dispositions in accordance with *p* being the case (Gendler, 2008). Consider the following paradigm cases (I have italicized in each case what that propositional content might be):

- When asked to choose between a set of identical consumer products, people *intuitively prefer the right-most product* (see *pantyhose study* in Nisbett & Wilson, 1977).
- “A ball and a bat cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?” Many people intuitively believe that *the ball costs 10 cents* (see *Cognitive Reflection Task* in Frederick, 2005).
- If someone asks me whether to wear a coat outside when its 20ºC, I know intuitively that *you don’t need to wear a coat when it’s 20ºC outside*. If someone asks me whether to wear a coat when it’s 293.15 Kelvin, my answer will be the same. But, this time, it is not the result of intuition (Gilovich, 2012).

Intuitions like these may result from fairly complex mental operations, that are deployed quickly, escape introspective awareness and in some cases contradict our rationally held, overt beliefs (Zajonc, 1980). For instance, an integrated network of our visual and linguistic systems enables a quick computation over the properties of electromagnetic radiation to derive its associated color term. This computation is unconscious and happens very quickly. In the moral domain, intuitionist views, like universal moral grammar, posit that core moral judgment is like the above intuitions: it exhibits propositional content, such as the ascription of deontic properties to actions under consideration (e.g., attributing *impermissibility*, to the target action of *breaking one's promise to a dead relative*. See Mikhail, 2007).

A second difference between affect and intuition is that:

Affect necessarily has a valence and a magnitude whereas intuitions do not.

We commonly talk of affective signals in terms of positive versus negative affect, behavior-inhibiting versus behavior-activating, approach-related versus avoidance-related; all of these pairs of terms reflect the basic insight that affective signals have the property of *valence*. Similarly, labels like blunted/reduced and heightened affect, demonstrate that we understand affect to have some *magnitude* as well. We can therefore differentiate thick emotional concepts along these two dimensions of their underlying affective states. For instance, *feeling pleased* has a positive valence, while *feeling upset* has a negative valence. Meanwhile, *feeling distraught* and *feeling annoyed* both have a negative valence, but are differentiated along the dimension of magnitude: *feeling distraught* has a greater magnitude than *feeling annoyed* has. (Notice that intuitions need not have these properties. With the exception perhaps of the "pantyhose" intuition[1], the above intuitions do not have valence and magnitude as constituent properties.) On a non-cognitivist account, moral judgments are more like these affective states; they do not carry propositional content but are necessarily linked to the evaluator's phenomenal experience of some sensation, either positive or negative and that can be stronger or weaker, in response to the consideration of the action.

1.4. Conclusion

I have reviewed some of the principal positions in the debate surrounding the psychology of moral judgment. Scientists from various disciplines – social psychology

[1] If one accepts my differentiation of affect and intuition, it might make sense to characterize this phenomenon as an affect-laden intuition.

(Haidt, 2001), but also cognitive neuroscience (Greene et al., 2001; Moll et al., 2002b), cultural anthropology (Shweder et al., 1997), and primatology (de Waal, 2009) to name a few – have examined morality as a natural phenomenon by employing their corresponding methodologies. Throughout the following chapters, I will provide my own interpretation of these findings and present some original experimental results with the goal of defending a novel, naturalistic theory of moral judgment.

In the following two chapters, I will review and critically discuss a variety of psychological evidence concerning moral judgment. Specifically, in Chapter 2, I will examine the putative role of reason in making moral judgments. In past decades, ample evidence on *automaticity* in decision-making led to the widespread view that many of our social judgments are conducted automatically, and uncontrolled by our conscious reasoning (Zajonc, 1980). Rather, reasoning is employed essentially for purposes of confabulation, i.e., to justify and explain our own conduct *ex post facto* (Fotopoulou, Conway, & Solms, 2007). This research program, which has since covered considerable ground in moral psychology, therefore casts doubt on the plausibility of cognitivism about moral judgment.

Just as recent evidence on decision-making and social cognition undermines the power of reason to govern behavior, it has equally vindicated the influence of emotion and affective processes more generally. So, in Chapter 3, I will try to differentiate some of the empirical claims that have been made surrounding the relationship between affective processes and moral judgment, and I will arbitrate between them. I will also aim, in the second half of the chapter, to characterize the psychological processes that causally produce these emotional responses, a question that has received markedly less attention. In essence, these two chapters together will serve as an exercise in analyzing past experimental studies in order to generate new hypotheses that shed light on the question of moral psychology. These hypotheses are then tested in Chapters 4 and 5.

In Chapter 4, I present a set of original, collaborative studies that examine people's normal aversion to immoral behavior. In these studies, we differentiate two putative kinds of affect in moral judgment: the aversion to the outcomes of the action (e.g., empathic concern) versus the intrinsic aversion to the action. We then ask whether the condemnation of harmful and disgusting behavior stems principally from either aversion. Our results draw support for the role of action aversion in moral judgments across both domains, and more broadly for the role of mental simulation in moral judgment.

Next, in Chapter 5, I present a set of studies that examine the influence of reason on moral attitudes. Here we find that individual differences in thinking style are linked with contrasting approaches to moral judgment. Specifically, the tendency to enjoy effortful thought and employ reasoning is linked with judging actions based on an assessment of the outcomes that they bring about. Meanwhile, the tendency to shy away from hard thinking and rely on gut feelings, known as an intuitive cognitive style, is associated with moralizing behavior on the basis of a consideration of the action itself. We also find that an emphasis on outcomes is associated with a morality centered on harm and fairness concerns, and an emphasis on actions was associated with a broader set of moral concerns including loyalty to the ingroup, respect for authority and purity-related norms.

Finally, in Chapter 6 I interpret the above findings and develop a new account of the human faculty for making moral judgments. Our findings provide evidence for two approaches to moral evaluation: One approach involves an evaluative simulation of the agent's behavior, and another involves an assessment of the outcomes befalling any victims and beneficiaries. These approaches can be construed in terms of a dual-process theory of moral judgment, congruent with several lines of research across a broader set of disciplines. First, evidence from neuroimaging studies indicates the double dissociation of two networks responsible for action versus outcome-based judgments. Second, computational models and behavioral studies in the learning algorithm literature document two approaches to decision-making that resemble action- and outcome-based decision processes. Third, evolutionary considerations suggest that some of our deeply-rooted aversions to actions may have provided an adaptive advantage to members of ancestral communities. So, the evaluative simulation approach serves as a heuristic that delivers the same decision as the outcome-based strategy in many contexts, but faster and with less cognitive effort. In Chapter 6, I argue that these different lines of evidence converge in support of the dual-process theory outlined above.

Chapter 2. Dual process theories and the role of reasoning in moral judgment

2.1. Introduction

A popular, but nevertheless contentious, framework for understanding human judgment and decision-making involves an appeal to a division between two fundamentally distinct types of mental processes. The origins of this widespread intuition – that human cognition is essentially of two opposed kinds – can be traced as far back as Ancient Greek discussions of the *parts of the soul* in Aristotle (1999) and Plato (1991), and located in modern works about the mind as the distinction between *reason and passion* in Hume (1739), or between *reason and the unconscious* in Freud (1990; see also James, 1890/1950). Freud’s (1990) dual theory of information processing distinguished between a primary system that is associative and unconscious, and a secondary system that is conscious and capable of rational thought. James (1890) regarded human reasoning also as either an experiential-associative type of thinking, or of an analytical-deliberative kind.

More recently, since the dawn of automaticity research in the cognitive sciences, numerous influential theories in psychology have assumed a fundamental differentiation between two systems in the mind (Evans & Over, 1996; Kahneman & Frederick, 2002; Shafir & LeBoeuf, 2002; Sloman, 1996; Stanovich & West, 2000). These systems have received diverse labels – heuristic versus analytic (Evans, 1989), associative versus rule-based (Sloman, 1996), automatic versus controlled (Stanovich & West, 1998) – and are often characterized by various other attributes, fast versus slow, intuitive versus reflective, high capacity versus low capacity, unconscious versus conscious, domain-specific versus domain-general, implicit versus explicit, and so on (see Evans & Stanovich, 2013). Distinctions of this sort have become the matter of much contemporary research on human reasoning and learning (Barbey & Sloman, 2007; Daw

& Shohamy, 2008; Dienes & Perner, 1999; Kahneman, 2011; Reber, 1993; Stanovich, 2011; Stanovich & West, 2000; Sun, Slusarz, & Terry, 2005; Sutton & Barto, 1999), especially in the domain of social cognition (Bargh & Ferguson, 2000; Chaiken & Trope, 1999; Epstein, 1994; Greenwald & Banaji, 1995; Kruglanski & Orehek, 2007; Smith & DeCoster, 2000; Wegner & Bargh, 1998).

Though the precise formulation of two systems has remained far from clear, it is understood that each of these distinctions differentiates two cognitive systems, and theories that posit the operation of two such systems are known as *dual process* or *dual systems* theories. This owes to certain core features of dual systems that these accounts, at least implicitly, accept. It is important, before entering a discussion about the proper characterization of moral psychology from a dual systems perspective, that we outline four of these central features (see Table 1):

Table 1. *Four central features of dual-process theories*.

Feature	*Description*	*References*
1) Parallel processing:	These two cognitive systems – System 1 and System 2 – can operate *in parallel*, i.e., simultaneously.	Epstein, 1994. Sloman, 1996. Stanovich & West, 2000.
2) *Computational load*:	System 1 is *associative*, whereas System 2 is *algorithmic*.	Epstein, 1994. James, 1890. Tversky & Kahneman, 1983.
3) *Temporal primacy*:	Operation of System 1 is *fast* and *unaware* whereas operation of the System 2 is *slow* and *controllable*.	Evans & Over, 1996. Freud, 1900. Tversky & Kahneman, 1983.

4) *Evolutionary primacy*:	System 1 has a longer evolutionary history than System 2.	Evans, 2003. Stanovich & West, 2000.

In the investigation of moral judgment specifically, a dual process framework has also garnered substantial popularity. Although not without its numerous detractors (Bartels, 2008, Nichols & Mallon, 2006, Mikhail, 2007), several theorists have either explicitly defended a dual-process view of moral judgment (Cushman, 2012; Greene, 2007; Greene & Haidt, 2002; Haidt, 2001), or profited from this kind of characterization of moral cognition (Baron, 1994; Sunstein, 2005).

In this chapter, I will review two highly influential dual-process accounts of moral judgment, Haidt's (2001) social intuitionist model and Greene's (2007) dual-process account of utilitarianism and deontology. Haidt and Greene agree that our moral judgments are often the result of System 1 processes: fast, unconscious and the result of a moral sense which evolved to facilitate life in small-scale, ancestral communities. This moral sense is partly shared, not just with our ancestors, but with non-human primates also (see de Waal, 2009), and has equipped us with a set of "hard-wired", spontaneous responses that readily influence our moral attitudes.

Yet the focus of this chapter is on a particular disagreement between Haidt and Greene, concerning the influence of System 2 processing on moral judgment. I will argue that this disagreement should be resolved by conceding that Greene's theory better accounts for the variety of data about the relationship between moral judgment and controlled cognition. Lastly, I will pose some problems for the characterization of controlled processing as rational, and elaborate a particular account of System 2 processing in moral judgment, according to which it involves the integration of rational and affective processes. Specifically, I will argue that System 2 moral judgments are composed of consciously-held beliefs, motivated by the engagement of affective processes.[2]

2.2. Haidt's social intuitionist model

In an early study of the condemnation of harmless moral violations, Haidt and colleagues (2000) examined participants' moral judgments about a case of harmless consensual incest:

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

This experiment serves as an exemplar of Haidt's (2001) dual-process theory of moral judgment, known as the *social intuitionist model*. A majority of the participants in this study judged this case to be wrong. Then, when prompted by the experimenter, appealed to some constructed reason citing a harmful outcome which was in fact precluded by the design of the vignette. Ultimately, participants clung onto the judgment that the behavior is wrong, even though they could not give an adequate justification for the judgment. This finding illustrates social intuitionism in a nutshell. According to social intuitionism, moral judgments are influenced by "a small set of intuitions that evolution has prepared the human mind to develop" (Haidt & Bjorklund, 2008, 181). These intuitions are "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" (ibid, 2008, 188).

[2] Notice, however, that I will not argue for a dual-process view here, a task which I leave for Chapter 6.

By contrast, the role of System 2 processing is (as demonstrated in this experiment) limited to constructing a verbal justification that supports one's prior moral intuition when faced with the social demand to do so (see also Nisbett & Wilson, 1977).

The social intuitionist model, though, stresses the influence of System 1 processes on moral judgment and behavior, while reason's influence on moral judgment is limited to cases of *reasoned persuasion*. That is, social intuitionism "gives moral reasoning a causal role in moral judgment, but only when reasoning runs through other people" (Haidt & Bjorklund, 2008, 193). Since justifications carry an affective valence, "reasoned persuasion works not by providing logically compelling arguments but by triggering new affectively valenced intuitions in the listener" (Haidt, 2001, 819).[3] In other words, the affective force of the justification elicits congruence from the listener, independently of the content of the justification. In this way, social intuitionism is indeed "social", by stressing the influences of interpersonal relationships, and of the verbal exchange of reasons, in reshaping one's moral intuitions.

Meanwhile, the direct influence of cognitive processing on moral judgment, according to social intuitionism, is hypothesized to be rare. Sometimes, *reasoned judgment* may take place when the initial System 1 intuition is weak, and/or System 2 processing capacity is high enough to counteract the intuition. Other times, *private reflection* may take place when various intuitions arise and the final judgment is determined "either by going with the strongest intuition or by allowing reason to choose among the alternatives on the basis of the conscious application of a rule or principle" (ibid).

[3] Reasoned persuasion should be contrasted with social persuasion, a distinct mechanism in Haidt's social intuitionist model, by which social interactions shape moral attitudes. In social persuasion, "the mere fact that friends, allies and acquaintances have made a moral judgment exerts a direct influence on others, even if no reasoned persuasion is used" (2001, 819). Therefore reasoned persuasion requires deliberation, whereas social persuasion does not.

2.3. Greene's dual process theory

A contrasting dual-process account of moral judgment has been put forth by Greene (2007) and colleagues (Greene et al., 2001, 2004). The basic insight of this theory is that a proper understanding of our evolved, moral minds ought to differentiate between evolutionarily ancient (or *personal*) moral violations that would have been familiar to our evolutionary ancestors – such as physical violence, or the perils of incest -- and evolutionarily new (or *impersonal*) ones that are unique to the modern environment – such as tax evasion and contamination by greenhouse gases. *Ex hypothesi*, an evolved moral sensibility would likely exhibit System 1 responses to personal violations but not impersonal violations. Therefore, while System 1 processes may guide moral judgments about evolutionarily ancient forms of harm, System 2 may exert a greater influence on judgments about evolutionarily modern kinds of moral violations.

This approach provides an attractive interpretation of diverse empirical data on moral judgment. For instance, it may explain the pattern of intuitions elicited by the famous trolley problem (Fischer & Ravizza, 1992; Foot, 1967; Thomson, 1985). Consider the following *switch* dilemma:

> A runaway trolley is racing down the tracks toward a group of five people who will be killed if it continues on its present course. However, you can save the five people by pulling a lever that will divert the trolley onto a different set of tracks. On this second set of tracks, there is only one person, who will be killed if you divert the train.
>
> Is it morally permissible to turn the trolley by pulling the switch, thus saving the five?

As is well-known, in this *switch* version of the dilemma people usually say that it would indeed be morally permissible. On Greene's view this is motivated by a simple (System 2) cost-benefit analysis. By contrast, consider the *footbridge* version:

> Once again, the trolley is headed for five people, and it will kill them if it continues on its present course. You are standing on a footbridge over the tracks, next to a large man. You can save the five people by pushing the man off the footbridge and into the path of the trolley, where his weight will stop the train in time to save the five people.
>
> Is it morally permissible to stop the trolley by pushing the man, thus saving the five?

People ordinarily judge that this would be morally wrong (Petrinovich, O'Neill, & Jorgensen, 1993). In this case, an evolutionarily ancient (System 1) aversion to forceful, intentional harm outweighs cost-benefit considerations. Greene gives an analogous explanation of the moralization of numerous sexual taboos, such as incest and homosexuality. The moral condemnation of these behaviors is argued to originate in evolutionarily ancient affective responses while more permissive attitudes are supported by controlled processing (Paxton, Ungar & Greene, 2012). Moreover, Greene and colleagues link these distinct neurocognitive systems to two broad traditions in the history of normative ethics – deontology and utilitarianism; specifically:

> The social-emotional responses that we've inherited from our primate ancestors (due, presumably, to some adaptive advantage they conferred), shaped and refined by culture bound experience, undergird the absolute prohibitions that are central to deontology. In contrast, the "moral calculus" that defines utilitarianism is made possible by more recently evolved structures in the frontal lobes that support abstract thinking and high-level cognitive control (Greene et al., 2004, 398).

2.4. The role of controlled processing

In sum, Haidt's and Greene's accounts share some core features: they both grant the primacy of System 1 processes in moral judgment, and they give an evolutionary account of these affect-laden intuitions. An obvious locus of disagreement between these views, however, concerns the role that each theory attributes to controlled processes of reasoning in shaping moral judgment. To recapitulate, on the social intuitionist view, controlled processes do not directly shape moral judgment (save in exceptional circumstances). We might term Haidt's view a *futility* view, according to which controlled cognition is largely futile in moral judgment. On Greene's view, controlled processes regularly influence moral judgment, specifically by promoting a calculation of aggregate welfare. This view might be labeled an *efficacy* view, i.e., where cognitive processes are indeed attributed a direct causal influence on moral judgment.

In what follows, I will bring prior evidence to bear in order to arbitrate between a futility and an efficacy view about the role of controlled processes in moral judgment. First, I will present an overview of the empirical literature and show that it draws support for an efficacy view. I will then articulate some potential responses on behalf of the futility view, and ultimately argue that these do not go through. Finally, I will distinguish a few versions of the efficacy view and defend a novel view which, though compatible with Greene's account, provides some much-needed detail to the coarsely defined role of controlled processing in moral judgment.

Numerous findings draw preliminary support for the efficacy view. These findings can be fruitfully categorized into two principal kinds. In one group of findings, *individual difference* studies demonstrate correlations between dispositional differences in the reliance on controlled processing and distinct moral judgment patterns (Bartels, 2008; Feltz & Cokely, 2008; see also Chapter 5). In another, *experimental* studies show that the experimental induction of controlled processing influences moral judgment (Paxton et al., 2012).

2.4.1. Individual difference studies

Primary evidence for the relationship between dispositional reliance on controlled processing and moral judgment comes from studies that index controlled processing through various, related measures of cognitive style.

For instance, Bartels (2008) employs a self-reported measure of thinking style, the Reflective-Experiential Inventory (REI; Epstein, Pacini, Denes-Raj, & Heier, 1996). On the REI, participants rate their level of agreement with statements like ‘‘I prefer to do something that challenges my thinking abilities rather than something that requires little thought” that suggest a deliberative thinking style, and others like ‘‘Using my gut feelings usually works well for me in figuring out problems in my life”, that suggest a tendency to rely on intuition. This study finds that participants who agree with statements like the former tend to endorse harm for the greater good while participants who agree with the latter category of statements tend to condemn harm for the greater good.

There are obvious limitations to a self-report paradigm. Most notably, on scales that measure desirable (or undesirable) personality traits, participants are often inclined to respond normatively rather than in a way that reflects them accurately. In addition, subjectivity in anchoring and interpreting the values on scales, render a self-report methodology limited at the least.

This problem can be addressed by employing *performance* measures rather than self-report measures of a given psychological trait. To this end, Frederick (2005), designed the Cognitive Reflection Test, a set of a few questions that have intuitively appealing yet evidently false answers. Consider the following example:

> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

People frequently give the first response that comes to mind, $.10, without thinking further and realizing that the bat would then have to cost $1.10, and therefore that the total cost would be $1.20 (instead of $1.10). Upon considering this problem more closely the correct answer is clearly $.05. The items on the CRT share this basic structure: An automatic and incorrect response springs to mind due perhaps to surface features of the problem, while the correct answer can be apprehended through further consideration. So, the CRT appears to serve as a valuable performance measure of participants' reliance on intuition versus reflection.

Two recent studies have examined the relationship between CRT scores and responding on trolley-type moral dilemmas (Feltz & Cokely, 2008; Paxton et al., 2012). These studies find that the more right answers a given participant provides, the more likely he/she is to endorse harming one to save many in personal dilemma contexts, like the footbridge dilemma.

Crucially, the dimension of cognitive style measures a personality trait that is somewhat independent of reasoning *ability*. Reasoning abilities are no guarantee against making the kind of judgment errors provoked by the CRT. In fact, in the original (2005) study, Frederick found that large numbers of highly select university students provided the intuitive but wrong answer. The prevalence of the erroneous answer indicates that the mind spontaneously recruits "thrifty" System 1 processes. Rather than measuring reasoning abilities, the CRT tracks participants' willingness or tendency to engage these reasoning (i.e., System 2) abilities in circumstances in which an intuitive (System 1) response is available. In this sense, cognitive *style* is the willingness or tendency to engage reflection to address a problem, and not the ability to do so.

However, endorsement of harm in footbridge-type dilemmas is associated with measures of cognitive *ability* also. Moore, Clark and Kane (2008) examined the relationship between *working memory capacity* (WMC) and moral judgment. WMC refers to the ability to hold multiple pieces of information in mind for short-term processing. As a measure of WMC, participants were asked to complete a processing task, such as verifying the meaningfulness of a sentence, the result a simple equation, judging whether a pattern is symmetrical while keeping in mind a working memory probe, i.e., an ordered list of letters or the location of several squares on the screen.

Participants' success rate in remembering the memory probes served as the measure of working memory capacity. Moore and colleagues (2008) found that participants with greater WMC were more likely to endorse harmful action for the greater good than were participants with lesser WMC, and this effect was selective for personal moral dilemmas. In addition, high WMC participants showed less variability in their responses to personal dilemmas than did low WMC participants, but this effect too was limited to the personal cases.

Convergent evidence for this claim derives from neuroimaging studies that examine brain activity during moral judgment tasks. Greene and colleagues (2004) found that, across subjects, endorsement of harm in personal dilemmas was associated with increased activity in the anterior DLPFC (BA 10), right inferior parietal lobe (BA 40) and in anterior regions of the posterior cingulate (BA 23/31). The anterior DLPFC and inferior parietal lobes are regularly observed to activate during working memory tasks and other comparable tasks requiring cognitive effort and/or abstract reasoning (Wager & Smith, 2003).

2.4.1.1. The reflective deontologist objection

An all-too-common, yet misguided, objection to these studies is predicated on the observation that numerous highly reflective individuals – including, influential moral philosophers such as Immanuel Kant – would disapprove of intentionally sacrificing one to save many. It is sometimes suggested that this observation challenges the proposed interpretation of the above studies, but this is not the case and I will explain why.

The objective of most naturalistic theories is to give a *general* account of moral judgment; i.e., to provide an account how most people (i.e., psychotypical humans) make moral judgments most of the time. Of course, depending on the specific methods by which data are collected certain subsets of this target sample will be excluded (e.g., in typical online studies, people who cannot read in English and people who do not have access to a computer are largely excluded). To the extent that these methods succeed in sampling randomly from the true population (of English speaking individuals with computer access), the resulting sample will include influential philosophers, reflective deontologists, intuitive utilitarians and so on roughly in the proportions in which they appear in the true population. And, as far as we can tell, when we do so, the result we obtain is the one reported above.

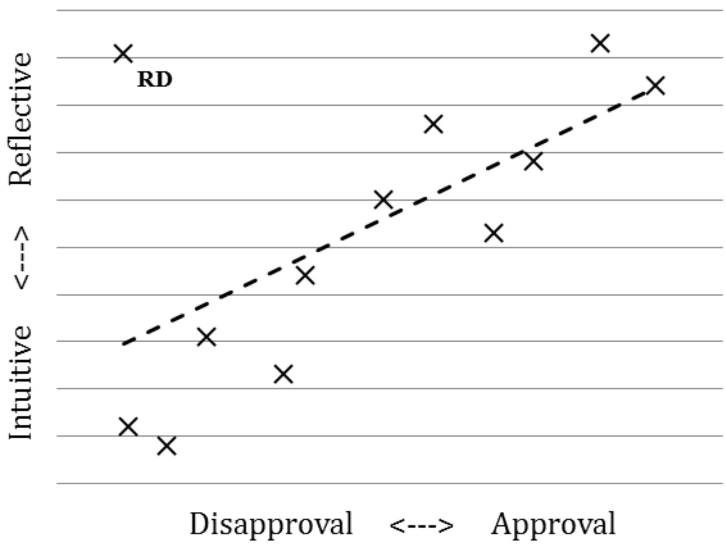


Figure 2. The reflective deontologist objection: A visual presentation.

Of course, the resulting trend captures only some proportion of the total variability in moral judgment. This means that knowing a certain individual’s cognitive style only determines a certain *range* within which the individual’s moral judgment lies with a certain *probability*. For outliers like Kant and other reflective individuals who disapprove of the welfare sacrifice (see the RD data point in Figure 2), this trend does a particularly poor job of predicting their moral judgment. But, this is to be expected; we can explain this by pointing to decades of scholarly reflection on morality as a likely influence on the moral judgments of deontological philosophers. So, it is important to recognize from the outset that the existence of outliers (or counterexamples), like Kant, does not undermine the observation of a probabilistic trend within a broader, target population.

2.4.1.2. The reshaping objection

Though in principle the above studies yield support for an efficacy view, there are numerous alternative explanations that one could give in line with a futility view. The following passage gives us a sense of how social intuitionists might reply to these data, in order to defend the futility view:

> Gut feelings say “no, don’t kill the child,” yet as soon as one leans toward making the “no” response, one must deal with the consequence that the choice leads to death for many people, including the baby. Greene’s fMRI data show that, in these difficult cases in particular, the dorsolateral prefrontal cortex is active, indicating “cooler” reasoning processes at work. But does a slow “yes” response indicate the victory of the sort of reasoning a philosopher would respect over dumb emotional processes? We think such cases are rather the paradigm of the sort of affective reasoning that James and Damasio described: there is indeed a conflict between potential responses, and additional areas of the brain become active to help resolve this conflict, but ultimately the person decides based on a feeling of rightness, rather than a deduction of some kind (Haidt & Bjorklund, 2008, 195).

This passage indicates a potential response social intuitionists might provide, which I will refer to as the *reshaping reply*. The reshaping reply is the following: When a dispositionally reflective person makes a permissive judgment about the footbridge dilemma, they do not – as Greene’s account would have it – override a forceful and deeply rooted condemnatory intuition with an exercise of cognitive control, and impose a permissive judgment generated by System 2 processing. Instead, over time the reflective individual has reshaped their very intuitions such that when they face a moral dilemma their automatic intuition is in fact a permissive one. This reshaping could, as Haidt suggests, happen in a dialectical context via socialization with other reflective thinkers. In other words, the social intuitionist interpretation of these results is that controlled cognition exerts an indirect effect on moral judgment, by facilitating a different intuition. In this way, the proximate cause of more permissive moral judgments about the footbridge dilemma is not a System 2 process but rather another System 1 intuition that has been revised over time by the tendency to engage in reflection, perhaps embedded in social, discursive practices.

This response thereby maintains that the moral judgments are intuitions, while granting (as the empirical evidence demands) that cognitive processing may lead over time to more permissive moral attitudes about these cases. (The question of *why* this is so I will leave for Chapter 6). For now, let us grant that the reshaping reply undermines the apparent implication of the correlational data, and look next at the experimental studies, which may turn out to be incompatible with holding the futility view.

2.4.2. Experimental studies

So-called *experimental* paradigms provide an alternate strategy for empirical testing, which in principle enables us to test whether System 2 processes causally influence moral judgment or instead follow in the wake of moral judgment-making (as is upheld by the futility view). With this methodology, initial equivalence is established between research participants composing more than one group, or *condition*. Thereafter, participants in the experimental condition(s) are subjected to a manipulation, while participants in the corresponding control condition are not. This manipulation should be designed to modify, boost or inhibit, certain psychological properties or dispositions. In our present case, the relevant psychological capacity is *the reliance on System 2 processing*. Next, participants complete the tasks of interest, containing the dependent variables which we seek to measure. If initial equivalence was established between the groups, and one group was subjected to a manipulation to which the other group was not, any measurable differences on the subsequent task are attributable to the causal influence of the manipulation. In our present case, manipulating affect prior to a moral judgment task – i.e., in order to boost reliance on affective cues – purportedly serves to examine the causal link from affect to moral judgment.

A recent study employed this approach, examining the influence of temporary boosts in reasoning processes on moral judgment (Paxton et al., 2012). Paxton and colleagues (2012) employed two manipulations aimed at promoting skepticism about one's intuitive judgments and fostering a reflective approach to moral judgment. In a first manipulation, participants completed the CRT prior to making moral judgments (versus after making moral judgments in the control group).

Paxton and colleagues observed that this manipulation led to increased endorsement of harm in personal moral dilemmas as predicted by Greene's dual process theory: participants who completed the CRT prior to the moral dilemmas judged the actions as more permissible than did participants who completed the CRT after the moral dilemmas. The suggestion here is that completion of the CRT induces more reflective responding in subsequent tasks, as participants have successfully engaged a System 2 response where a System 1 response was available (see also Pinillos, Smith, Nair, Marchetto, & Mun, 2011).

This influence of controlled cognitive processes is not limited to moral judgments about harm. It is found also in judgments about violations of the purity domain. Paxton and colleagues (2012) manipulated controlled processing before participants judged Haidt's harmless incest case. Half of the participants read a strong debunking argument for the evolutionary basis of the disgust response to incest, while the other half read a weak argument. In each group, half of the participants were requested to respond immediately and the other half were required to delay their response by at least two minutes to promote reflection, yielding four groups of participants. Paxton and colleagues observed that giving people enough time to ponder the strong argument selectively made participants' judgments about the case more lenient. (That is, neither viewing the strong argument alone independent of reflection time, nor being given time to reflect independent of argument strength influenced moral judgments significantly.) This once again suggests that System 2 processing – slow deliberation over consciously held information – influences moral judgment.

Together, the two manipulations led participants to revise their intuitive, System 1 response, and engage cognitive processes to provide a counterintuitive, System 2 response instead. At face value, these experimental results support the efficacy view about the role of controlled processing.

2.4.2.1. Social intuitionist replies

Here again the social intuitionist will likely issue a version of the reply outlined earlier, i.e., that controlled cognition indirectly shapes moral judgment by triggering an alternative intuition. But, even if social intuitionists will argue that intuition ultimately shapes moral judgment, it seems that here they must concede that controlled cognition is involved in something beyond the dialectical exchange of *post hoc* justifications. After all, the influence of these manipulations on moral judgment cannot be explained by differences in the opportunity for social and dialectical exchange. It is not obvious that viewing an argument on-screen counts as a 'social' exchange, but let's suppose it does. Still, reasoned persuasion – as described in the social intuitionist model – cannot explain this pattern of results. Recall that, when reasoned persuasion takes place, it is the affective valence of *A*'s (the computer's) reasoning that influence *B*'s (the participant's) moral judgment. But both the strong and weak arguments in the above experiment exhibit the same attitude of permissibility towards consensual, non-reproductive incest. So, if this effect were attributable to reasoned persuasion, we would expect the strong and weak arguments to influence participants equally.

It seems instead that what is influencing participants' judgments is precisely the rationality of the argument. Social intuitionists will at least have to concede that controlled processing is implicated in something like *suppressing the predominant intuition*, i.e., concerning the wrongness of pushing the man off the footbridge in one case, and of consensual incest in the other. This is compatible with the futility view, just as long as the proximate cause of utilitarian judgment is still *some* System 1 process. Still, this departs significantly from the standard, social intuitionist account of the role of controlled processing.

A further question in this social intuitionist interpretation of the above data concerns the origin of the utilitarian intuition. Supposing utilitarian attitudes were to stem from a System 1 process, social intuitionists would still have to offer some account of what kind of intuitions these judgments stem from, while respecting the four core features of System 1 intuition, which I presented at the beginning of this chapter. Let us differentiate two potential replies, which I will then consider in sequence:

1. Reshaping reply: The participant did not have a prior utilitarian intuition. The manipulations of controlled processing (the argument and deliberation time, and the CRT) originated a *novel* utilitarian intuition.
2. Framing reply: The participant had a utilitarian intuition prior to the manipulation. The manipulations of controlled processing (the argument and deliberation time, and the CRT) bring out this *dormant* utilitarian intuition.

The first response, a variant of the reshaping reply discusser earlier, seems less compatible with the results of the experimental studies than it was with the individual difference studies. This is due at least to a couple of considerations. The first is that the reshaping response (to these causal data) violates one of the core features of System 1 processing. This feature of dual-process theories, about the evolutionary-ontogenetic primacy of intuition states that System 1 processes originate early in phylogeny and/or ontogeny, by comparison to System 2 processes (Evans, 2003). This precludes labeling as a System 1 response any psychological state that both originated and manifested within the average span of a testing session. A second consideration against the reshaping response is that social persuasion, as understood by Haidt (2001), cannot account for the reshaping of participants' intuitions. On several counts, it appears that the reshaping response to the causal data fails.

Next let's consider the framing reply. According to this view, these experimental manipulations serve to frame moral questions in such a light that brings "dormant" (but pre-existing) utilitarian intuitions to the forefront. This explanation seems more plausible than the reshaping reply. But what kind of intuition would this dormant utilitarian intuition be?

To answer this question, let's borrow Haidt's own five-part taxonomy of moral intuitions (Haidt & Graham, 2007). According to Haidt and Graham (2007), human moral intuitions belong to five, distinct moral foundations. That is, moral intuitions can be fruitfully categorized into five clusters, and both evolutionary (Haidt & Graham, 2007) and statistical (Graham, Haidt & Nosek, 2009) considerations support this classification. According to moral foundations theorists, the human mind is predisposed to acquire moral intuitions corresponding to each of these five foundations, as observed in numerous moral codes across cultures (Haidt & Joseph, 2004). The foundations of *harm* and *fairness* encompass norms that proscribe harm to others and unjust behavior, and are suggested to relate to the evolution of the mammalian attachment system, empathy and mind-reading, and the development of reciprocal altruism. The *loyalty* and *authority* foundations are composed of norms that establish societal order and maintain intra-group relations e.g., norms about loyalty to superiors, patriotic duties, respect for elders, as well as govern attitudes towards the outgroup. These foundations are proposed to be closely linked to the evolution of small-scale, hunter-gatherer and tribal communities and their hierarchical structures. Finally, the *purity* foundation contains a wide range of norms concerning food, hygiene, proper use of the body, including sexual behavior, and religious mandates about transcending carnal and animal impulses, and cultivating a spiritual sense. These norms evolved originally to deter the ingestion of contaminant substances, but now serve a broader purpose of establishing standards of decency and propriety. So, which moral foundation might inspire the intuition to condone killing one to save many? What about the intuition to demoralize consensual incest?

A good candidate, at least with respect to our first question, would be *fairness*. This intuition, thought to have a relatively long evolutionary history (Haidt & Graham, 2007; Rand, Greene & Nowak, 2012), could lead us to consider it morally right to value each of the lives of the five distal victims just as we value the life of the one proximal victim, with the likely consequence of promoting a utilitarian judgment about this case.[4] It is less clear, by contrast, what kind of intuition would yield the *de*-moralization of sexual taboos, such as consensual incest. But, even if we suppose that there is a latent intuition in support of de-moralization. Still, the framing reply fails to account for the temporal primacy of System 1 processes. Temporal primacy, remember, claims that System 1 responses are elicited quickly by the presence of the relevant stimuli. Quite the contrary, in Paxton and colleagues' experiment, permissive moral judgments required a certain lag. In this sense, it remains somewhat unclear whether these judgments – the endorsement of personal harm in footbridge-type dilemmas, and the demoralization of sexual taboos – might be properly characterized as System 1 responses.

2.4.3. Summing things up

To sum things up, at face value these data lend credence to the efficacy view. I outlined some interpretations of these data that a social intuitionist might provide. First,

[4] Indeed, some unpublished evidence shows that fairness may undergird this moral view. We conducted a multiple regression analysis predicting moral judgments about footbridge style cases with harm and fairness scores as independent predictors. This analysis revealed distinct effects of both scores, such that harm increased condemnation and fairness *decreased* condemnation of footbridge dilemma cases. This suggests that fairness concerns may undergird the endorsement of personal harm for the greater good.

with respect to the individual difference studies, social intuitionists might argue that dispositionally reflective people have reshaped their intuitions over time through dialectical exchange with like-minded individuals; a reply which holds some ground. Second, with respect to the experimental studies, Paxton and colleagues' causal manipulations frame moral issues in ways that support certain, fairness-related, intuitions which are otherwise dormant and favor utilitarian responding.

One might find these alternative explanations more or less convincing, but in any case they are elaborate, or even tedious, by comparison to Greene's parsimonious explanation. In addition, social intuitionists must explain a looming and mysterious connection between controlled processing and these alternate utilitarian intuitions: Why do reflective thinkers tend to demoralize personal harm and purity violations? And why do these experimental manipulations favor permissive moral attitudes over the spontaneous, condemnatory responses? There are no hints in the social intuitionist literature about what this connection might be.

Though we have not yet arbitrated between the efficacy and futility views, in passing we have demonstrated that the role of System 2 processing is likely greater than social intuitionism has characterized it as being. The influence of controlled cognition on moral judgment is not limited to that of providing justifications in the context of dialectical exchange, but includes also an involvement in the suppression of predominant moral intuitions (such as those serving to condemn utilitarian trade-offs and purity violations). Having established this much, in the next section I will present some data that weighs in on our primary discussion: i.e., the debate between the efficacy and futility views.

2.4.4. Are moral judgments algorithmic?

As we saw earlier, multiple features of utilitarian moral judgment – its latency, its susceptibility to argument, and so on – suggest that it may be primarily a controlled, System 1 response rather than an intuitive, System 2 response. However, these considerations are not decisive since we can imagine plausible interpretations of the relevant findings from a social intuitionist perspective. I will now review some evidence from a recent neuroimaging study (Shenhav & Greene, 2010) that rather convincingly demonstrates that utilitarian moral judgment violates an important postulate of System 1 processing: it is algorithmic (or compositional), rather than associative. That is, rather than a sudden flash of valenced affect, utilitarian moral judgment is the product of a mental operation that consists of several steps.

In the aforementioned neuroimaging study, Shenhav and Greene (2010) presented participants with a series of moral dilemmas. For each moral dilemma, the probability and magnitude of the consequences were manipulated. For instance, in one scenario, the agent is on a rescue boat headed to save a drowning man when she receives another emergency call alerting her of another wreckage involving a larger number M of passengers. There is another rescue boat that is near the larger wreckage and which will succeed in saving the M passengers with probability P. Participants are requested to judge whether it would be morally acceptable or unacceptable to head towards the larger wreckage and leave the single drowning man to die. The experimenters systematically manipulated M and P, and recorded participants' moral judgments as well as their neural activity during trials.

The neural data revealed distinct brain regions that tracked the *probability of loss of life*, $1 – P$ (in the right anterior insula), the *magnitude of harm* or number of lives in the larger wreckage, M (in the central insula, dorsal striatum, and anterior and posterior cingulate cortices), and the *expected value* of heading towards the larger wreckage, $M - M \times P - 1$ (in the ventral striatum, ventromedial prefrontal and medial orbitofrontal cortices). In addition, participants' moral judgments (about whether it would be permissible to head toward the larger wreckage and leave the single victim to drown) were sensitive to the expected value of the action. This study therefore demonstrates that distinct neural mechanisms encode the probability, magnitude and expected value of behavioral options in a moral context (and that, as shown in previous studies too, this expected value calculus influences participants' moral judgments about these cases). Notice that this evidence strongly suggests that utilitarian moral judgment implicates an algorithmic, System 2 process, rather than some (perhaps fairness-related) intuitive flash. That is, participants are retrieving the information about probability, retrieving the information about the magnitude of harm, and deriving the product of these two values.

These data are incompatible with the futility view. It is hard to imagine a convincing social intuitionist interpretation of these results that would serve to put the futility view back on the table. The social intuitionist would have to argue against the cumulative evidence that these distinct pieces of information in utilitarian calculus are being tracked by different neural regions *and* that this mental operation influences subsequent moral judgments. The claim that, even in these cases, moral judgment is a System 1 response is likely to garner meager support. Of course, social intuitionists might cast doubt on the external validity of Shenhav and Greene's experimental design altogether. In other words, they might grant that this study documents the occurrence in the lab of a calculated utilitarian reasoning, but question the representativeness of these moral dilemmas. That is, they might claim that this experimental task is too dissimilar to normal moral decision-making contexts for the elicited judgments to represent ordinary moral judgment. This I think will ultimately fail also. People ordinarily make moral and non-moral decisions that are consistent with the maximization of subjective value. Though indeed the probability and magnitude information were made explicit and salient in this experiment, there is little reason to believe that – when this information must be sought, and perhaps these values estimated – fundamentally different cognitive processes are recruited instead (which don't involve the value calculus based on probability and magnitude). In sum, we ought to retain Greene's more parsimonious interpretation of these data, according to which these data demonstrate that the demoralization of personal harm and purity violations depend on an algorithmic mental operation, i.e., on controlled processes of reasoning.

2.5. Is welfare calculus affectively motivated?

In this last section, I will advance one last claim about the nature and operation of System 2 in moral judgment. I will argue that to characterize the dual systems, and their corresponding roles in moral judgment, as "rational" and "emotional" is misguided. The evidence reviewed above may explain why this first-pass distinction gained some traction in early theoretical discussions: i.e., both the reasoned debunking of System 1 responses and the algorithmic processing involved in constructing utilitarian judgment can be thought of as "rational" processes. However, in this section, I will demonstrate that moral judgments cannot be composed of rational processes alone. I will then defend an alternate view according to which both System 1 and System 2 approaches to moral judgment are rooted in affective processes. On this account, the controlled system accomplishes utilitarian moral judgment by computing an algorithmic calculus, *motivated by an affective valuation of patient welfare*. Specifically, I will argue that determining the magnitude (but not the probability) of an outcome recruits an affective psychological mechanism. Therefore, without this affective basis (which I will seek to explain in the remainder of this chapter), one could not make a normal, System 2 moral judgment.[5]

Recall that, in Shenhav and Greene's study, utilitarian judgment was made up of two neurocognitive components – *probability* and *magnitude* – and their product, *value*.

[5] As a side note, it is not clear whether the claim I will put forth here, that affect is involved in utilitarian moral judgment, is at odds with Greene's (2007) view.

Making a utilitarian judgment involves aggregating the outcomes of an action *A*, each of which has a probability P_i, magnitude M_i and an expected value of $M_i \times P_i$. Consider the mental states that might undergird each of these components. Assessing the expected probability P_1 of outcome O_1 is straightforwardly a matter of having a *belief* about the conditional probability of O_1 on *A* (i.e., about the likelihood that the outcome will ensue if the action is carried out). It is easy to see that this capacity does not require an affective basis (but only a belief about the probabilistic, causal relations that obtain). By contrast, in order to assess the moral magnitude of an outcome, my claim in this section is that having the relevant beliefs about an outcome *O* (for instance, concerning the number of times *O* will happen) does not suffice. Encoding the magnitude of an outcome requires in addition a *valenced attitude* towards the outcome itself.

A couple of core requirements of utilitarian decision-making, I will argue, substantiate this point. Consider, first of all, the *ability to compare* outcomes, specifically, outcomes that are different in kind. This characteristic of ordinary utilitarian judgment is somewhat obfuscated in the experimental literature where moral judgment tasks often present unrealistically quantifiable trade-offs: sacrificing one life vs. saving five lives, and so on. But in many real life contexts, the utilitarian comparison of outcomes is not as straightforward, because one’s behavioral options involve *qualitatively* (and not merely quantitatively) distinct outcomes. So, utilitarian decision-making requires also the ability to compute the expected value of qualitatively distinct outcomes, and in particular, to determine their relative magnitude. For instance, it seems patently obvious that the magnitude of *breaking a leg* exceeds the magnitude of *losing a finger nail*. But how psychologically do we carry out this comparison? Some mechanism enables us to determine (rather rapidly) that breaking a leg entails a greater degree of harm, a longer term incapacitation, etc., than does losing a finger nail, and correspondingly in utilitarian decision-making we prefer the *losing a finger nail* outcome to the *breaking a leg* outcome (supposing, let's say, that the expected probability of both outcomes is the same).

Second, consider a related capacity that plausibly requires an affective underpinning, i.e., the capacity to value others' welfare *at all*. Making controlled moral judgments implicates the capacity to *value* or *prefer* utilitarian outcomes and not merely to *identify* them as such. In crude terms, one might say that a controlled moral judgment does not merely yield the belief that "saving a life *is more than* saving no lives" but rather the preference for saving a life, i.e., "saving a life *is (morally) better than* saving no lives". This points towards the presence some underlying non-cognitive attitude towards others' welfare. By analogy, consider the non-moral choice between a brick and a bottle of apple juice. In addition to having beliefs about these alternatives, in order to *prefer* one option over the other one must have a certain non-cognitive attitude, for instance, thirst. The evaluator's thirst grants value to the apple juice, but not to the brick. Similarly, there must be some non-cognitive attitude, brought to bear in utilitarian decision-making that grants (negative) value to the outcome "the old lady broke her back", but not to "the cardboard box broke". With this non-cognitive attitude (but not without it), the evaluator may have the preference for the latter outcome over the former outcome. This non-cognitive moral attitude would grant value *simpliciter* to the welfare of others (and, when aggregated across the multiple, probabilistic outcomes of an action, yields a preference for utilitarian courses of action).

In sum, I have shown that ascribing a moral magnitude to the outcomes of an action requires some affective mechanism. Specifically, I have outlined two abilities that are commonly recruited by everyday utilitarian decision-making, and which are hard to account for in terms of rational or deliberative processes alone: (i) the comparison of qualitatively distinct outcomes, and (ii) the valuation of others' welfare *simpliciter*. These core requirements point towards the involvement of some non-cognitive moral attitude that motivates the welfare calculus in utilitarian decision-making.

2.5.1. Simulation theory of empathy

The above considerations point tentatively towards the existence of a non-cognitive attitude towards others' welfare that enables the ascription of valence and relative magnitude to distinct outcomes. In this section, I will propose a candidate affective mechanism that could yield this non-cognitive attitude: *empathy*. This psychological capacity was first examined closely by the Scottish Enlightenment philosophers (Hume, 1739; Smith, 1759/2011), and in recent years, has received notable empirical substantiation in the neurosciences and cognitive psychology, under the rubric of the *simulation theory of empathy* (for a review see Gallese, Keysers & Rizzolatti, 2004). In reviewing the pertinent literature, I will demonstrate how empathy induced via simulation of the patients' perspective might compose the affective basis of System 2 moral judgment, and that it can facilitate the valuation and comparison of distinct moral outcomes.

It is well known that several Scottish Enlightenment philosophers argued that "sympathy" provides the foundation of morality. Consider, for instance, the following passages from Hume and Smith:

> We may begin with considering a-new the nature and force of sympathy. The minds of all men are similar in their feelings and operations; nor can anyone be actuated by any affection, of which all others are not, in some degree, susceptible. As in strings equally wound up, the motion of one communicates itself to the rest; so all the affections readily pass from one person to another, and beget correspondent movements in every human creature. When I see the effects of passion in the voice and gesture of any person, my mind immediately passes from these effects to their causes, and forms such a lively idea of the passion, as is presently converted into the passion itself. In like manner, when I perceive the causes of any emotion, my mind is conveyed to the effects, and is actuated with a like emotion (Hume, 1739, 3.3.1).

> By the imagination we place ourselves in his situation, we conceive ourselves enduring all the same torments, we enter as it were into his body, and become in

> some measure the same person with him, and thence form some idea of his sensations, and even feel something which, though weaker in degree, is not altogether unlike them (Smith, 1759, 1.1.1).

The basic picture on their view was that, whether triggered by imagination or observation, the evaluator could simulate the patient's experience, perhaps including her beliefs, desires and motivational set, and in so doing, experience a similar emotion or hedonic state to that experienced by the patient.

In the contemporary cognitive sciences, this philosophical account has received remarkable validation. The foremost advance in the empirical investigation of simulation-based empathy was the original discovery of single *mirror neurons* in macaque monkeys (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992), and analogous networks later in humans (Iacoboni et al., 1999). These mirror neurons fire both for own execution of the action, and for observation of a similar action in another individual. Later work expanded from the recognition of motor actions to explore the role of mirror neurons in the social ascription of emotional states to others. It is this body of work which brings to bear closely on the capacity for empathy discussed by Hume and Smith. Several studies have now demonstrated the existence of mirror neurons for the recognition of emotions (via facial expressions) and pain, activated indifferently when participants re requested to either imitate or merely observe emotional facial expressions (see e.g., Carr et al., 2003).

Mirror neuron activation has also been observed in the experience of pain: participants who observe others' pain exhibit activation in networks that are linked to the first-hand experience of pain (Avenanti, Paluello, Bufalari, & Aglioti, 2006; Botvinick et al., 2005; Morrison et al., 2004; Singer et al., 2004). A number of other studies, using magnetoencephalography and functional MRI, have since demonstrated that empathy for pain involves also the somatosensory cortex (Cheng, Yang, Lin, Lee, & Decety, 2008; Lamm, Nusbaum, Meltzoff, & Decety, 2007; Moriguchi el al., 2007;

Ogino et al., 2007). Critically, an empathic response, facilitated by mirror neuronal activation, is observed not only following *observation* of pain and emotional states, but also following the mere *imagination* of another's experience of pain (Singer et al., 2004). Precisely because observation and imagination of pain and emotional states in others activate neural networks for own experience of the perceived state, these representations of others' welfare activate somatic and autonomic responses to own experience of pain, and thereby initiate certain appropriate, pro-social behavioral responses (Preston & de Waal, 2002). Now let's examine how this psychological faculty might fulfill our desiderata; i.e., how simulation may yield some non-cognitive attitude towards others' welfare, and facilitate the attribution of a moral valence and magnitude to specific outcomes.

First, the representations of others' hedonic states inherit the relevant properties from first-hand instantiation of comparable hedonic or affective states. This will likely include the association of a valence (e.g., positive) to recognizable emotional states (e.g., excitement). So, the third-party attribution inherits the valence associated to the first-person knowledge of the particular affective state. Second, mental simulation yields a certain non-cognitive attitude towards other's welfare for similar reasons. By processing the instantiation of others' emotional and hedonic states through mirror neuronal circuitry, one is motivated to respond in similar ways (though perhaps not in the same degree) than one responds to own states, resulting in the prosocial valuation of others' welfare.

What about the ability to compare outcomes by ascribing relative magnitudes? Does simulation grant *this* ability? Indeed observation of emotional states of different magnitude yields differences in the strength of mirror neuronal activation and ensuing autonomic response (Harrison, Singer, Rotshtein, Dolan, & Critchley, 2006). It is known, for instance, that pupils are decreased in size with the experience in sadness.

Harrison and colleagues therefore presented images of sad faces while manipulating the size of the stimuli’s pupils. Pupil size was found to influence the magnitude of the ascribed sadness, the strength of somatic and autonomic response (as measured by activation in the amygdala), and was mimicked by the participant while viewing the stimuli. This evidence demonstrates that differences in the magnitude of perceived harm are essentially tracked by differences in the evaluator’s empathic response. This suggests a capacity to ascribe a magnitude to an outcome under consideration, which can thereafter be compared to another outcome (Saarela et al., 2007).

In this section, we reviewed some evidence demonstrating how empathy, via the mental simulation of the patient’s affective state, could undergird the capacities required by moral judgment which cannot be accounted for by rational and deliberative processes alone. The observation or imagination of another’s experience elicits, through the activation of shared systems, a comparable emotional state in the evaluator. By understanding another’s emotional states as neural instantiations of one’s own comparable emotional states, this mechanism is capable of (i) yielding a non-cognitive moral attitude towards others’ welfare, which (ii) correlates with the perceived magnitude or severity of the patient’s projected state. In this way, the empathic response derived from simulation of the patient’s experience enables the ascription of a valence and relative magnitude to different moral outcomes under consideration.

2.5.2. Piecing it together

The simulation theory of empathy explains how simulating the victim’s or beneficiary’s perspective might induce an empathic response which grants a valence and a relative magnitude to the outcome under consideration. Now we are in a position to evaluate how this affective mechanism complements Greene’s neurocognitive account of the welfare calculus involved in utilitarian moral judgment. The relative magnitude of the outcome, along with a belief about the probability of the outcome, together yield the outcome's expected value. The expected value of each behavioral option is compared, and a moral judgment is made by endorsing the behavioral option with the greater expected value for patient welfare. In what follows, I will articulate and address two obvious weaknesses of the proposed account of patient simulation.

2.5.2.1. Psychological overdemandingness

One might worry that this account of moral judgment is simply too contrived to reflect any aspect of the operation of folk moral judgment: the complexity of the mental operations stipulated by this account exceeds what might plausibly be attributed to ordinary moral cognition. Of course, in trolley-type moral dilemmas, where a very limited number of outcomes are stipulated by the context, the cognitive operations demanded by my proposed account might not overdemanding. In these circumstances, the agent's behavioral options (pushing a fat man off the footbridge vs. not) and the corresponding morally relevant outcomes (the death of the one vs. the death of the five) are severely restricted. By contrast, applying utilitarian moral decision-making in real-life contexts (e.g., which career path to take, or which of a variety of products to buy) involves the choice between multiple behavioral options, and the morally relevant outcomes extend beyond the immediate consequences of the action. So for these contexts, we might question whether a mechanism that entails the consideration of *all* relevant outcomes is a plausible feature of ordinary moral cognition.

Indeed it is rather unlikely that controlled moral judgment involves the consideration of all morally relevant outcomes in these more realistic contexts. But the account I have defended is not committed to such a claim. There is no reason to suppose that the exercise of controlled cognition requires the consideration of *all* morally relevant outcomes. Instead it is more appropriate to suppose that the range of morally relevant outcomes that figure into the aggregate welfare calculus may be influenced by situational and individual differences, such the evaluator's own cognitive style. Indeed, as we saw before, reflective individuals tend to make more utilitarian moral judgments. On the account I am developing we can explain why this is: they are better able to, and more willing to, hold a greater number of morally relevant outcomes in mind, with their corresponding probabilities and magnitudes, and aggregate them to estimate the expected moral value of distinct behavioral options. This taxing mental exercise is one that reflective individuals are more likely to accomplish. In sum, controlled processing in moral judgment *enables* optimal utilitarian moral judgment, but it is important to see that it does not *necessitate* utilitarian moral judgment. This account can still explain the existing correlational data without being psychologically overdemanding.

2.6. Conclusion

The account I have put forth bears some resemblance to Greene’s dual process theory. First of all, I grant Greene’s efficacy view is closer to the truth than Haidt’s futility view: controlled processing plays a role in favoring utilitarian moral judgment. However, Greene’s account understates the necessity for an affective underpinning to utilitarian judgment. I.e., controlled cognition must be supplemented by a valuation mechanism, in order to yield an evaluative attitude. In this way, both System 1 and System 2 are ultimately grounded in affective processes. Controlled cognition has multiple roles to play in moral judgment: first of all, to suppress the System 1 response, and second, to perform the welfare-maximizing calculus. However, as I said, these are altogether insufficient to constitute a moral judgment without a mechanism of valuation, which in utilitarian judgment is facilitated by empathy for other’s welfare, psychologically enacted via simulation.

Chapter 3. Mental simulation and the spontaneous evaluation of moral events

3.1. Introduction

While the influence of reason has been challenged, the role of affect in human moral competence has been vindicated by a wealth of empirical evidence. Dozens of studies highlight the pervasive role of affect in folk judgments about the moral permissibility of actions (Greene et al., 2001; Inbar, Pizarro & Bloom, 2009a; Schnall et al., 2008), and the moral character (Inbar, Pizarro & Cushman, 2012) and responsibility (Nichols & Knobe, 2007) of agents. In this chapter, I will highlight two unresolved questions concerning the formation of spontaneous moral judgments. Before moving forward, I will briefly unpack what is meant by *spontaneous moral judgment* throughout this chapter. I will be employing this term to refer to the average moral judgment made by psychotypical individuals, and which springs relatively quickly to mind in the wake of a morally evaluable event.

The precise characterization of the role of affect in moral judgment has been subject to intense theoretical debate. In the first part of this chapter, I will take part in this debate by restating some of the different claims that have been made on this matter, and evaluating them in light of the diverse empirical studies that bear on this question. I will argue that the view according to which affect merely results from moral judgment-making is untenable.

Whereas the question about the psychological character of moral judgments has drawn great attention (i.e., Are moral judgments beliefs, intuitions, emotions?), very little progress has been made in characterizing the antecedent psychological processes which give rise to them, i.e., those processes taking place between the presentation of moral stimuli and the resulting moral evaluation. In the second half of this chapter, my aim will be to push forth in the characterization of these processes. I will claim, first, that several popular accounts of moral judgment assume that moral judgment must derive from some process of categorizing actions on the basis of their properties. I challenge this assumption by examining a range of studies that, I argue, are better explained by appealing to processes of mental *simulation*. In passing, this defense of simulation-based processes over categorization-based processes in turn speaks to our debate about the character of moral judgment, advocating a strong causal role for affective processes.

Process

Stimuli → Mental state

Figure 3. The structure of moral evaluation.

Therefore, my general strategy in this paper will be to work backwards in the causal chain (see Figure 3). First I will examine the psychological constitution of moral judgments (the *mental state*) in Section 2. Next, in Sections 3 and 4, I will look closer at the processes which yield this mental state (the *process*).

3.2. Characterizing the relationship of affect to moral judgment

In the last two decades, ample experimental evidence has been produced by empirical studies in moral psychology which demonstrate an involvement of affective processes in the formation of spontaneous moral judgment. Throughout this section, I will discuss much of that body of evidence in order to adjudicate between three popular and distinct claims about the relationship between affect and moral judgment. Before getting started, I will differentiate three such claims that vary in the *strength* of the relationship they posit between affect and moral judgment:

1. Affect as *constitutive*. The activation (or induction) of a certain affective state causally produces moral judgment.

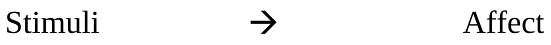


Figure 4. Affect as constitutive.

This is the strongest claim, and versions of it are endorsed by Prinz (2006), and Haidt (2001). According to this view, the processing of moral stimuli yields an affective response, positively or negatively valenced, which constitutes the moral judgment (see Figure 4). Therefore the induction of an affective state is capable of interfering with moral judgment-making processes (e.g., causing the misattribution of moral valence to a non-moral action). This sort of view is best paired with an *emotivist* or *sentimentalist* account of moral judgment.

2. Affect as *modulator*. Moral affect influences moral judgment. Specifically, moral judgments are constituted of intuition, but the experience of moral affect during moral judgment-making process influences the severity of moral judgment.

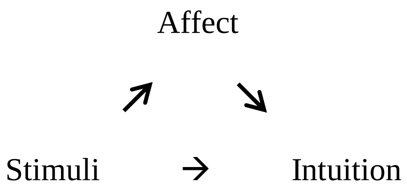


Figure 5. Affect as modulator.

According to this view, affect does not play a necessary causal role in moral judgment. It does however influence the moral decision-making process, acting as an “amplifier/dampener” of moral intuition (see Figure 5). This view predicts that the induction of moral affect cannot by itself yield a valenced moral judgment. But in conjunction with a moral intuition, affect can modulate the severity of ensuing moral judgment. So, this claim seems to speak in favor of a moral *intuitionism*, and has been upheld by a number of moral psychologists (Horberg, Oveis, & Keltner, 2011; Huebner et al., 2009; Nichols, 2002).

3. <u>Affect as *epiphenomenon.*</u> Moral affect does not influence moral judgment. Rather, moral affect results from processes of moral judgment-making.

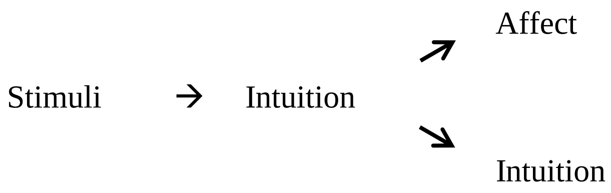


Figure 6a. Affect as epiphenomenon I.

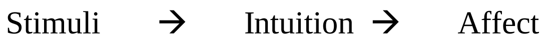


Figure 6b. Affect as epiphenomenon II.

Notice that there is a fundamental difference between the epiphenomenal view, and the constitutive and modulator views. The epiphenomenal view grants no causal role to affect in shaping moral judgment: on this view, affective responses merely follow from the process of moral judgment-making (as in Figure 6b), or from the presentation of moral stimuli in a manner independent of the moral judgment processes (as in Figure 6a). Epiphenomenal claims of this sort fit most naturally in support of moral *intuitionism*, and versions of them are found in the work of Mikhail (2011).

Having differentiated three claims about the relationship between moral affect and judgment, in the remainder of this section I will examine the existing evidence in order to arbitrate between these views, and ultimately, to shed light on the question whether the core mental state in spontaneous moral judgment is an intuitive or an affective mental state.

3.2.1. Correlational studies

Early empirical studies show that distinct moral emotions, each with its corresponding facial expressions, are associated to different kinds of moral concerns (Rozin et al., 1999b). According to Rozin and colleagues' (1999b) *CAD triad hypothesis*, anger, contempt, and disgust are typically and cross-culturally elicited by violations of autonomy (individual rights violations), community (violation of communal codes, including hierarchy), and divinity (violations of purity-sanctity) correspondingly.[6] Participants in the United States and Japan were shown a series of descriptions of situations, such as a child saying dirty words to his parents (*community*), a man coming home drunk and beating his wife (*autonomy*), or touching a corpse (*divinity*). They were asked to pair each situation with the appropriate emotional response, by selecting either a facial expression or an emotion term corresponding to contempt, anger, or disgust. In a further experiment, participants read these situations and were asked to make appropriate facial expressions in response. Facial expressions were videotaped and coded according to movements that characterize each emotion (e.g., a one-sided smirk for contempt, a lowered brow for anger, or a wrinkled nose for disgust). Results on these studies provided clear support for the CAD triad hypothesis, indicating that unique emotions correspond to distinct moral domains.

The association between trait disgust and violations of divinity (or purity) has been examined in greater depth in subsequent studies (Horberg, Oveis, Keltner, & Cohen, 2009; Inbar et al., 2009a). People vary in the strength of their reactions to disgusting situations, like seeing a human hand preserved in a jar during science class or drinking from a used but disinfected toilet bowl. Participants' sensitivity to disgust has been shown to predict judgments about purity-related issues, like incest, homosexual sex (Inbar et al., 2009a), such that participants who are more prone to feelings of disgust tend to condemn these behaviors more harshly. This relationship is also observed with participants' *implicit* judgments about male homosexuality, for example (Inbar, Pizarro, Knobe & Bloom, 2009b). Inbar and colleagues tested a sample of typically liberal college students to discover whether participants' implicit attitudes towards male homosexuality might be at odds with their explicit avowals. A long history of social

[6] Incidentally, this proposal is an important precursor of the moral foundations theory I discussed in Chapter 2.

psychological studies demonstrates that numerous biases and prejudices underlie participants' outwardly egalitarian commitments with respect to women, the elderly, racial minorities, and so on (Greenwald & Banaji, 1995). Indeed Inbar and colleagues found that, even among participants who overtly would express tolerance for male homosexuality, implicit measures of their attitudes revealed a negative bias. But, more importantly for our purposes, this bias was proportionate to participants' level of trait disgust, i.e., greater disgust sensitivity correlated with greater implicit prejudice and condemnation of homosexuality.

Neuroscientific research further demonstrates the involvement of affective systems during the presentation of morally valenced stimuli (Moll, Eslinger, & de Oliveira-Souza, 2001, Moll, de Oliveira-Souza, Bramati, & Grafman, 2002a, Moll et al., 2002b) and during the formation of moral judgments (Greene et al., 2001, 2004). Moll and colleagues (2001, 2002a, 2002b) report heightened affect and activation in emotional networks in the brain during the contemplation of moral stimuli.[7]

In a first study, subjects heard a set of moral and non-moral statements while in the fMRI scanner (Moll et al., 2001). Participants pondered the meaning of the sentence and judge whether they found each statement right or wrong. (The Portuguese terms "certo" and "errado" are fully ambiguous between moral rightness/wrongness and truth/falsity.) Relative to non-moral claims, consideration of claims with moral content produced the most robust activity bilaterally in the frontal pole (BA 10/46) and the medial frontal gyrus (BA 9; see also Moll et al., 2002). A subsequent study by Moll and colleagues found similar neural responses to pictures with moral content, such as physical assaults, or poor abandoned children (Moll et al., 2002b), relative to control,

[7] These studies provide also convergent behavioral evidence of the correlation between affective processes and moral judgment: Participants judged moral statements, like "The boy stole his mother's savings", to be more emotionally charged than non-moral statements, like "Desserts make you fat"; and their moral judgments of the actions were predicted by their ratings of emotionality for both right and wrong actions.

non-moral images. Across all three studies, activation of the medial frontal gyrus (BA9) – observed during attention to subjective emotional states (Lane et al., 1997) and, in particular, of unpleasant emotions (Lane et al., 1999; Damasio et al., 2000) – was preferentially engaged during consideration of moral stimuli.

Greene and colleagues' (2001, 2004) contrasting approach examined the brain activity of individuals while making moral judgments about a series of moral dilemmas. Some of the dilemmas were *impersonal*, where agents choose to save a greater number of people and sacrifice someone as a side-effect of so doing (as in the switch case). The other half were *personal* dilemmas, in which the agent actively harms a victim in order to save a greater number of people (as in the footbridge case). In line with prior data, they found increased participants condemn the personal trade-offs more severely than the impersonal trade-offs (Petrinovich, O'Neill, & Jorgensen, 1993). Additionally, judgments of personal moral dilemmas elicited greater activation in the medial prefrontal cortex (BA 9/10), posterior cingulate gyrus (BA 31/7), and bilateral superior temporal sulcus/inferior parietal lobe (BA 39), than did judgments of impersonal dilemmas. These brain regions are associated with affective processing, suggesting that affect is involved in the condemnation of intentional sacrifices in personal moral dilemmas.

3.2.1.1.Non-cognitivist interpretation

Greene's own non-cognitivist interpretation of these data is fairly straightforward: The moral mind is such that the consideration of violent forms of harm and/or violations of sexual taboos, as seen in the incest case, elicit System 1 affective responses directly. These 'hard-wired' affective responses constitute the moral judgments we make.[8] This view therefore predicts that participants who experience

[8] As we saw, on Greene's account, the presence of heightened affect during moral judgment tasks is the result of evolutionary pressures that shaped our psychology to respond quickly and effectively to certain upclose moral violations. These affective

stronger affective responses will condemn the moral violations more heavily, while participants who experience weaker affective responses condemn them less. Differences in affect – whether indexed by neural, physiological, or self-report measures – are associated to differences in moral judgment since these affective responses themselves constitute the moral judgments.

3.2.1.2. Intuitionist interpretation

In response, universal moral grammar theorists, such as Mikhail (2007) and Huebner (2008) who defend a view of moral judgment as constituted by non-affective intuition, have been eager to remind us that the methodologies employed in these correlational experiments do not warrant a causal claim about the involvement of affect. Whether behavioral, physiological or neuroscientific in kind, this collection of studies "provides only correlational data, showing that emotions are associated with moral judgments. Such data (on their own) can never be used to infer causality" (Huebner et al., 2009, 3). In addition, the fMRI studies, "because of the poor temporal resolution of neuroimaging, cannot be used to assess when emotions have a role or whether they are constitutive of moral concepts. In summary, the mere activity of neural circuits classically associated with emotion in processing moral scenarios fails to distinguish between the claim that (i) emotions are integral to moral computation and (ii) emotions result from these computations" (Huebner et al., 2009, 3-4). In sum, these studies do not arbitrate between the above views; they are compatible with a constitutive, modulator or epiphenomenal role for affect.

Mikhail's explanation of Greene's neural data rests on the assumption of intuitive, and plausibly innate, concepts in the domains of morality and law. Mikhail points out that Greene and the scientific community at large have neglected to see that:

responses played some adaptive role in dissuading the performance of these immoral behaviors.

> …all of the actions described by these [trolley-type] vignettes are well-known crimes or torts. […]By contrast, only five of the 19 cases in Greene’s “impersonal” condition are batteries, and only one of these batteries is purposeful. The basic cleavage he identified in the brain was not Kant versus Mill, but purposeful battery, rape, and murder, on the one hand, and a disorderly grab bag of theft crimes, regulatory crimes, torts against non-personal interests, and risk–risk tradeoffs, on the other. Moreover, his finding that the MPFC, PCC, STS, and amygdala are recruited for judgment tasks involving purposeful battery, rape, and murder does not undermine the traditional rationalist thesis that moral precepts are engraved in the mind. To the contrary, Greene’s evidence largely supports that thesis. […] Naturally, violent crimes and torts are more emotionally engaging than insider trading or environmental risk analysis, but it does not follow that emotion “constitutes” or “drives” the judgment that the former acts are wrong. Rather, what drive these intuitions are the unconscious computations that characterize these acts as battery, rape, or murder in the first place (Mikhail, 2011, 294).

So, for Mikhail, the personal/impersonal distinction differentiates grave crimes featuring murder or purposeful battery from an assortment of lesser crimes. An intuitive knowledge of moral and legal norms grants us the fundamental abilities to assign differential blame to these kinds of cases (just as our intuitive knowledge of linguistic norms, allows us to construct, understand and evaluate the grammaticality of sentences we have never heard). This innate knowledge contains concepts, such as “intentional battery”, “murder” and so on, which are cemented in our legal and moral lexicon. Therefore, on this intuitionist account, the role of affect – discovered in the trolley problem studies, but also in judgments about simple harmful actions – is limited to that of epiphenomenon; affect merely results from this intuitive parsing of actions according to innate moral and legal principles.

So, as it stands, these explanations are on equal footing: The aforementioned correlational evidence is compatible with either the constitutive and modulator views, latent in Greene and colleagues’ (2001, 2007) interpretation of these studies, or the epiphenomenal view, championed by Mikhail’s (2000, 2011) competing explanation.

3.2.2. Experimental studies

Several studies in moral psychology have employed experimental designs to manipulate the reliance on affective processing. For instance, in the examination of harm violations, Valdesolo and DeSteno (2006) presented participants with a comedic video clip as an induction of positive affect. Participants viewed a funny *Saturday Night Live* clip (in the experimental group) or an affectively neutral video clip (in the control group) and then responded to versions of the switch and footbridge dilemmas. The researchers reasoned as follows: If (as Greene's view posits) people condemn the action of pushing the man in front of the trolley *due to* a negative affective response, then the dose of positive affect induced by watching a short comedic sketch should counteract the negative response that drives moral condemnation, thereby making people's judgments less condemnatory. By contrast, if negative affect is a mere epiphenomenon, as Mikhail argues in the passage above, this manipulation should drive no difference in moral judgment. In line with Greene's view, the researchers found that people who had watched the funny video tended to endorse pushing the man in front of the trolley more often than did participants who had watched the neutral video clip. Similar results have been achieved with neurochemical manipulations of affective state (Crockett et al., 2010), where an increase in levels of serotonin – a neurotransmitter implicated in behavioral inhibition – was shown to lead to greater condemnation of welfare trade-offs in personal contexts.

Similar effects have been demonstrated in the condemnation of purity domain issues. Schnall and colleagues' (2008) subjected their participants to a variety of manipulations aimed at enhancing the experience of disgust – e.g., sitting at a messy desk or smelling a noxious odor during the testing session – and found that participants made more severe moral judgments than did controls. Similarly, the induction of disgust through hypnosis influences moral judgment (Wheatley & Haidt, 2005). Participants who were hypnotized to feel disgust in response to a neutral word like “often” then judged moral transgressions as more wrong in vignettes containing the hypnotically targeted word. Across a range of stimuli, experimental inductions of affect are shown to influence subsequent moral judgment tasks, indicating a causal role for affect.

3.2.2.1. Rejecting the epiphenomenal view

Intuitionists have offered a couple of distinct replies to refute the causal inference that these data seemingly warrant, acknowledging the influence of affective processes on moral judgment while retaining the core tenet of intuitionism. For instance, Huebner and colleagues argue that:

> …emotion could modify the inputs into distinctively moral circuits rather than modulating the operation of these moral circuits themselves. Thus, although asking subjects to evaluate a moral question triggers the process of moral evaluation, the negative emotional state yields a more severe moral judgment because of an increased focus on the ‘antecedently’ morally salient features of the scenario (2009, 3).

This response basically amounts to an endorsement of the modulator view; i.e., affect is not the core component of spontaneous moral judgment, but it can (and does, as demonstrated in these studies) exert an influence on moral judgment, primarily by modulating its severity. A distinct response offered by intuitionists characterizes the induction of affect as instigating the formation of a competing judgment:

> …although emotion yields ‘practical’ judgments, it is unclear that this warrants treating emotion as constitutive of ‘moral’ judgments. Perhaps moral cognition can be interfered with by introducing distracting emotional stimuli. However, *because disgust functions practically to help us avoid toxic, infectious or contaminating substances, it could generate interruptive judgments that could compete with moral cognition for attentional resources*. Existent data fail to address the plausible hypothesis that the apparent modulation of moral judgments by emotion is an artifact of the redeployment of attentional resources (Huebner et al., 2009, 3, italics added).

On this interpretation, disgust interferes with moral judgment processes by instigating the operation of a distinct class of judgments, according to Huebner and colleagues, ‘practical’ judgments. That is to say, when heightened disgust is activated, participants are engaged in the cognitive task of determining whether "any toxic, infectious or contaminant substances" are present, which should be avoided. Although Huebner and colleagues do not make this clear, the potential disgust elicitors could be the prime – i.e., the fart spray, or the dirty desk – or the moral stimuli itself – i.e., the case of consensual incest, bestiality, and so on – or both. In either case, the manipulation prompts participants to form the practical judgment to avoid the disgust-eliciting, potentially infectious or contaminant substance or practice. In sum, this account advocates a version of the modulator view as well, i.e., affective processes modulate moral judgment, but do not constitute an essential part of it.

3.2.3. In search of intuition

Whereas these experimental studies persuasively rule out the epiphenomenal view, they are compatible with both the constitutive view, defended by Greene (2007), and the modulator view, defended by Huebner and colleagues (2009). At this point, we may choose to accept a sort of stalemate between the constitutive and modulator views in terms of their explanatory power. And yet, if we bring considerations of theoretical parsimony to bear, it seems patently clear that the modulator view suffers an important disadvantage. While the ontology that is posited by a non-cognitivist account is tractable across numerous empirical studies, the intuitionist ontology requires the existence of some additional mental state, the *intuition*, for which no experimental evidence has been produced. . The above collection of studies, whether correlational or experimental, operationalize *affective states* via psychometric scale completion (e.g., disgust sensitivity), neural activation patterns (e.g. principally in the amygdala and VMPFC), or physiological signals (e.g., skin conductance response). So non-cognitivist explanations, like Greene and colleagues', can straightforwardly identify the indexed affective processes as the precise psychological states that *constitute* participants' moral judgments. Contrarily, intuitionists explain this evidence in every case by positing the existence and operation of a distinct, non-affective intuition that is nowhere tractable in the experimental literature. In simple terms, we might ask rhetorically: Where is the evidence for the *non-affective intuition*? In this respect, the constitutive view holds the advantage of theoretical parsimony over the modulator view.

3.3. Further back in the causal chain

So far in this chapter we have focused on the relationship between affective processes and moral judgment, in order to determine the psychological character of moral judgment. In the remaining sections, I will take a step back in the causal chain and examine the nature of the psychological *processes* deployed in the interpretation of moral stimuli. In other words, what are the cognitive operations that take place between the presentation of moral stimuli and the resulting affective response? Whereas a vibrant discussion has centered on determining the precise psychological character of moral judgment, relatively less progress has been made in characterizing the antecedent, cognitive processes which yield our moral judgments. In addition, our present discussion will provide some purchase on the emotivism-intuitionism debate which has been the focus of this chapter thus far. This is because, in characterizing the cognitive processes that yield the pattern of spontaneous moral judgments evinced in the experimental literature, we will arbitrate between the constitutive and modulator views.

3.3.1. Categorization/computation view

Numerous intuitionist accounts presuppose that moral intuitions are derived through some process of *computation* over, or *categorization* on the basis of constituent properties of moral actions. This view is rather common in moral psychology, appearing in Nichols and Mallon (2006), Bartels (2008), and Mikhail (2007). Nichols and Mallon (2006) argue that the *normative* properties of actions – that is whether they violate any normative rules – in conjunction with an affective response (of disgust, or empathic concern) jointly determine the action’s moral status. A related view, owing to Mikhail (2000), argues that moral judgment is the product of an operation over the *causal* and *intentional* properties of an action.

The case for intuitive categorization or computation has rested heavily on a number of studies concerning the so-called *means/side-effect distinction*:

> M/SE: Harm as a means to an end is morally worse than equivalent harm resulting as a side effect of achieving an end.

M/SE is observed routinely in folk moral judgments both in hypothetical dilemma contexts and in contemporary ethical debates. This principle might account, for instance, for the difference in people’s intuitions concerning killing, in active euthanasia, versus letting die, in passive euthanasia (Foot, 1967). Similarly, in the experimental literature, we observe that actions involving harm as a means, e.g., the footbridge case, are judged to be forbidden, while actions that involve harm as a side-effect, e.g., the switch case, are judged to be relatively more permissible. This pattern of intuitions arises rather universally and is normally deployed without the evaluator’s own introspective knowledge of the principle itself (Cushman, Young & Hauser, 2006; Royzman & Baron, 2002). For these reasons, M/SE has stood as a clear exemplar of an unconscious moral principle that is intuitive, and plausibly has some innate basis.

Categorization theorists have taken this principle as evidence that the moral status of an action is determined through a computation that depends on the action’s intentional properties: namely, target moral actions are categorized as permissible if they involve un-intentional harm, and forbidden if they involve intentional harm, in line with the M/SE principle. Nichols and Mallon, for instance, refer to this capacity in terms of “identifying rule violations”:

> [We] suggest that judgments on the footbridge cases are guided by affect-backed rules: our all-in judgment to footbridge-style cases is a product of both rules and emotions. […] Typically, then, when a person judges an action as all-in

> impermissible despite its having favorable outcomes this depends on both emotional activation and on thinking that a rule has been violated (2006, 540).

On their view, moral violations therefore depend on an automatic categorization that depends on two components, one affective – whether harm ensues – and one intuitive – whether a rule has been violated (intentionally). The moral status of actions, on this view, depends in essence on a computation over these two properties of actions (i) whether they elicit an affective response, and (ii) whether they involve the intentional violation of a rule (see Table 2). Note that this view can account also for the moral-conventional distinction: i.e., moral and conventional offenses alike involve the intentional violation of a rule, but moral offenses alone result in harm (Nichols, 2002). Both features are needed to produce a judgment of moral condemnation.

Table 2. *Two-component computation à la Nichols.*

	Rule violation:	**NO Rule violation:**
Victim distress:	*morally forbidden* (e.g., footbridge dilemma, hitting a child)	*morally permissible* (e.g., switch dilemma)
NO Victim distress:	*morally permissible* (e.g., wearing pajamas to the opera)	

Similarly, Mikhail's *universal grammar theory* (2000, 2007) accounts for the deployment of the M/SE distinction in terms of a categorization process, akin to that of the grammatical parsing of sentences. Consider, for example, the structural representations of the footbridge and switch dilemmas in Figure 7:

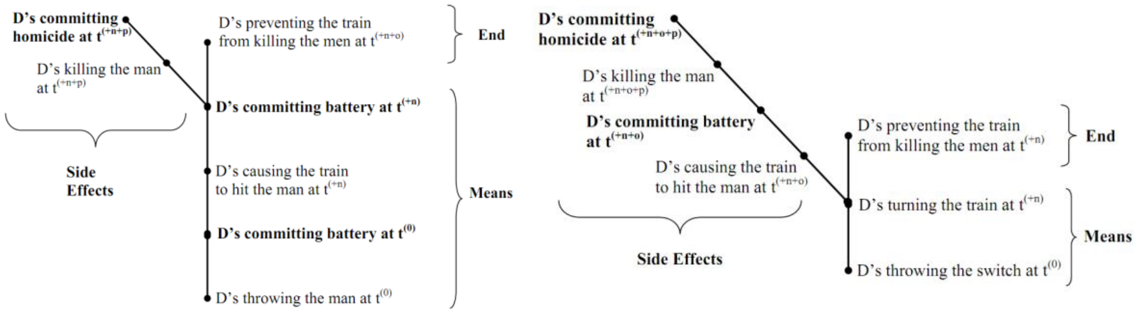

D's committing homicide at t^(+n+p)
D's killing the man at t^(+n+p)
Side Effects
D's preventing the train from killing the men at t^(+n+o)
End
D's committing battery at t^(+n)
D's causing the train to hit the man at t^(+n)
D's committing battery at t^(0)
D's throwing the man at t^(0)
Means
D's committing homicide at t^(+n+o+p)
D's killing the man at t^(+n+o+p)
D's committing battery at t^(+n+o)
D's causing the train to hit the man at t^(+n+o)
Side Effects
D's preventing the train from killing the men at t^(+n)
End
D's turning the train at t^(+n)
D's throwing the switch at t^(0)
Means

Figure 7. Mikhail's structural representations of (left) footbridge and (right) switch dilemmas.[9]

In each representation, we observe at the base the agent's original action. On the central branch we can see the agent's intentional behavior, while on diagonal branches we observe the side-effects of the agent's behavior. At the top of each central branch, we observe the agent's ultimate *goal*: i.e., to save the five workers on the tracks. Along each branch, we have a series of temporally distinct nodes that complete the description of the behavior. On Mikhail's view, target actions are categorized and ascribed deontic properties (i.e., forbidden, permissible, obligatory) on the basis of this complex structural representation of their temporal, causal and intentional properties.

In our present case, a difference in the representation of an action's intentional properties specifically gives rise to the M/SE distinction. In both cases, the agent commits battery and homicide against the proximal victim. However, the footbridge case involves battery as *a means to* saving the five lives, while in the switch case battery is *a side-effect of* saving the five lives (Mikhail, 2007). This is evinced in our descriptions of events using causal and intentional language: consider the linguistic expressions "*D* caused the train to hit the man *in order to* save the five workers" and "*D* saved the five workers *by* causing the train to hit the man". Notice that these statements are appropriate descriptions of the footbridge case, but fairly inappropriate descriptions of the switch case.

[9] Diagrams extracted from Mikhail (2007).

Mikhail's view therefore shares a fundamental perspective with Nichols & Mallon's. They both stress the unconscious computation over a few orthogonal factors. Some factors involve the detection of a harmful outcome, while other actors involve the detection of intention. In both accounts, the computation over these distinct components occurs quickly and unconsciously (even if the information may be consciously accessible *ex post facto*). Spontaneous moral judgment of the target action is dependent on the result of this computation.

In the remainder of this chapter, I will argue for an alternative to the categorization view. I will begin by presenting three sets of studies, concerning people's moral judgment patterns, that challenge the assumption that the central cognitive process in moral evaluation is one of intuitive categorization:

1. *Motor system*: Moral judgments are sensitive also to motor features of the agent's action (Greene et al., 2009).
2. *Sensory systems:* Moral judgment implicates sensory networks, including *visual* (Amit & Greene, 2012) and *olfactory* (Inbar, Pizarro & Bloom, 2012; Schnall et al., 2008) systems.
3. *Projection:* Moral judgment involves projecting one's *own* values (Bartels, 2008; Lieberman & Lobel, 2012) and non-moral preferences in third-party moral judgment.

I will argue that these studies do not fit squarely with the categorization view, and throughout the following section, I will present a fundamentally distinct view of the cognitive processing of moral stimuli that better accounts for these data.

3.3.2. The involvement of sensory and motor systems

Evidence suggests that both sensory and motor systems are involved in the formation of spontaneous moral judgment. In particular, moral judgments about personal moral dilemmas and purity violations are susceptible to manipulations that engage sensory and motor systems.

3.3.2.1. Motor system

In a 2009 paper, Greene and colleagues argued that the cognitive processes by which we normally condemn *intentional* harm, proscribe also the use of *motor force* in causing harm. This conclusion is based on the results of two experiments that explore the established difference in moral judgment between personal and impersonal moral dilemmas, in search of an added feature that drives condemnation of the personal cases. As I just discussed, one such feature is captured by the M/SE: the involvement of harm as a *means* versus as a *side-effect* of saving the five workers. However, the dissimilarity between the two cases is not exhausted by this distinction. It is also true, for example, that in the personal case the victim is *closer* to the agent than in the impersonal case, and that the agent comes into *contact* with the victim only in the personal case. So, it seems reasonable to ask: Which (if any) of these other differences is an active ingredient in the distinction between personal and impersonal harm?

In order to test this, in a first experiment Greene and colleagues constructed several candidate *personal*ness factors – spatial proximity, physical contact, and personal force – while holding the intention factor fixed across conditions (i.e., all conditions involved harm as a means to saving the five). In particular, the comparison involves the four conditions listed in Table 3:

Table 3. *Personal-ness factors and vignettes in Greene et al. (2009).*

	Physical contact	**Personal force**	**Spatial proximity**
Standard footbridge	✓	✓	✓
Footbridge pole	✗	✓	✓
Footbridge switch	✗	✗	✓
Remote footbridge	✗	✗	✗

Greene and colleagues found no significant difference between the Standard Footbridge and Footbridge Pole versions, indicating that physical contact did not drive a difference in moral judgment. Similarly, no difference was observed between Footbridge Switch and Remote Footbridge dilemmas, indicating that spatial proximity was not a morally relevant feature either, driving no difference in participants' moral judgments. By contrast, this study did reveal a difference between participants' moral judgments of dilemmas that featured personal force and dilemmas that did not. Therefore Greene and colleagues concluded that *personal force*, whether the agent's muscular force was involved in harming the victim, influenced participants' spontaneous moral judgments.

In a follow-up experiment, Greene and colleagues manipulated both the personal force and intention (i.e., means vs. side-effect) factors in moral dilemmas. Half the participants viewed a means case, and the other half viewed a side-effect case. In each of those groups, half the participants viewed a case involving the application of personal force and the other half viewed a case without personal force. The results indicated that participants drew the M/SE distinction only when personal force was present, and not when it was absent; so that when the agent harmed the proximal victim as a means to saving the five and by using his/her personal force, moral condemnation was greatest.

Together these results show that a personal force factor influences moral judgment (a result not predicted by the categorization theories), in conjunction with the intention factor, accounting for the differentiation between personal and impersonal harms. Greene and colleagues consider the significance of this interaction, drawing implications for the cognitive processing of moral stimuli:

> In a general sense, this suggests a mechanism of moral judgment that is a species of embodied cognition. One natural source of such embodied goal representations is a system of action planning that coordinates the application of personal force to objects to achieve goal-states for those specific objects. A putative sub-system of moral judgment, monitoring such action plans, might operate by rejecting any plan that entails harm as a goal-state to be achieved through the direct application of personal force. […] At a more general level, the present study strongly suggests that our sense of an action's moral wrongness is tethered to its more basic motor properties, and specifically that the intention factor is intimately bound up with our sensitivity to personal force (Greene et al., 2009, 370).

The suggestion here is that, rather than an intuitive categorization of the action, the relevant process by which spontaneous moral judgments are formed is a kind of 'embodied' cognition. On this view, bodily structures might be engaged in the cognitive processes subserving moral judgment.

Of course it is straightforwardly possible to account for this finding from within a categorization-based account by merely adding one more factor to the computation process, i.e., one that is sensitive to whether the action involves personal force or not. And it is, in principle, not necessary for this to require 'enactment' through bodily structures, as predicated by embodied cognition theorists. In sum, the evidence does not arbitrate between *a mental representation* and *a motor system enactment* of the action's motor properties (the second of which is Greene and colleagues' hasty interpretation).

3.3.2.3. Sensory systems

Spontaneous moral judgment seemingly involves also sensory neural systems, demonstrated by at least two sets of studies documenting the role of the visual (Amit & Greene, 2012) and olfactory (Inbar et al., 2012) systems in processes of spontaneous moral judgment. Amit and Greene (2012) conducted three experiments to examine the role of visual versus verbal cognition on moral judgment. In a first experiment, their approach was to develop a measure of differences in visual versus verbal thinking style. The premise here is that some people are better at thinking visually, while others are better with language. The task required viewing a target item along with two probe items, and selecting the probe item that is most similar to the target item. For instance, a certain visual trial might require determining whether a checkered red circle or a plain blue star is more similar to a checkered blue star. The corresponding verbal trial would involve columns of words (i.e., "checkered", "red", and "circle") instead of the depicted figures. Relative accuracy on the visual versus verbal portions of the matching tasks yielded an index of visual versus verbal thinking style. The researchers then contrasted this index of participants' thinking style with their moral judgment of trolley-type dilemmas. The results revealed that participants with a visual thinking style tended to condemn personal harm in trolley-type dilemmas more harshly than did participants with a verbal thinking style.

In a second experiment, Amit and Greene (2012) examined the causal role of visual and verbal working memory. Participants were asked to complete a *2-back* working memory task of either visual or verbal memory, interleaved with a moral judgment task. Participants viewed a series of shapes (in the visual condition) or shape names (in the verbal condition), and were required to indicate by means of a button press whether each item was identical to the item presented two items earlier. This time, the visual interference, relative to verbal interference and no interference, yielded reduced condemnation of personal moral dilemmas – plausibly because the visual system is implicated in processes of moral judgment. The authors argue that the target behavior is spontaneously visualized and this process gives rise to moral condemnation. So, when the visual system was occupied with a distracter task, resulting moral judgments tended to be more lenient. Participants' self-report confirmed this hypothesis: in the personal condition, participants were more likely to report seeing the proximal victim in their "mind's eye" than in the impersonal condition. These results therefore suggest a specific role for the visual system in the processing of moral stimuli.

We find convergent evidence for the role of sensory systems in processes of moral evaluation in studies that examine participants' judgments about the domain of purity (Schnall et al., 2008; Inbar et al., 2012). Studies have demonstrated that the induction of a bad smell (through the application of fart spray in the testing room) influences attitudes towards sibling marriage and sibling incest (Schnall et al., 2008) and homosexual men (Inbar et al., 2012), making them more negative by comparison to the control group in a neutral-smelling testing room.

Can a categorization view account for these different findings? As I suggested, the extant categorization views can account for the Greene and colleagues' (2009) finding without appealing to the involvement of motor systems –i.e., by merely adding a factor to the computation process, which is sensitive to whether the action involves personal force. However, the involvement of visual and olfactory systems (in Amit and Greene, 2012; Inbar et al., 2012; Schnall et al., 2008) seems more certain. This is because the corresponding studies directly engage visual and olfactory systems, through relatively specific manipulations of the content of these sensory systems. Therefore, the attempt to interpret these findings as evidence for categorization will fail.

3.3.4. Projection in moral judgment

It has frequently been noted that people exhibit the partly unconscious tendency to attribute one's own characteristics, goals and flaws to others (Freud, 1957; Holmes, 1978). For instance, people tend to overestimate the extent to which people share their beliefs and worldview (see *false consensus effect*, in Ross et al., 1977). In one study, participants were given the hypothetical choice between writing a group paper at the end of a class, or individual papers. Those who reported preferring to write a group paper estimated that more of their class would prefer the same than those who would prefer to write individual papers (and *vice versa*). Similar effects have been observed in the moral domain, on hot-button moral issues in United States policy, such as capital punishment and affirmative action (Ross et al., 1977). People who support affirmative action (/capital punishment) believe that support for affirmative action (/capital punishment) is more widespread than do people who oppose it, and the same is true *mutatis mutandis* for people who oppose these policies.

Projection might explain numerous other patterns of spontaneous moral judgment, for instance, the condemnation of third-party violations of one's own *protected values* (Bartels, 2008). Protected values are those that we think of as absolute or unquantifiable, such that they cannot be traded for anything else (Ritov & Baron, 1999). Whereas human life is more universally a protected value, other values such as the natural environment, animal life, religious symbols or works of art are much more culturally relative; i.e., they are clearly not seen as protected in all cultures or by all individuals.

Bartels (2008) sought to examine the influence of protected status on moral decision-making. The experiment first evaluated whether participants grant protected status to a set of twenty values (e.g., birds, children, dolphins, the poor, trees). Participants then viewed moral dilemmas involving trade-offs with these values. For example:

> A flash flood has changed the water levels upstream from a dam on a nearby river. Scientists estimate that 20 species of fish upstream from the dam are threatened with extinction. David is considering opening the dam, which will save these species, but some species downstream will become extinct because of the changing water level. Because this flood has rapidly changed water levels, a decision must be made quickly, and the government's options are severely constrained. David wants to save the fish species upstream. He first calculates that opening the dam will kill 16 species downstream. Knowing that doing so will kill many fish, he chooses to open the dam.

Participants then made moral judgments about the agent’s behavior. Results indicated that participants tended to judge trade-offs involving their *own* protected values as morally worse than did participants for whom those values were not protected.

This finding may appear rather trivial. We can explain it rather straightforwardly by appealing to *principled moralization*: that is, people have certain moral principles by which they judge others’ behavior (and their own). As a result of development, socialization or whatever, we adopt a series of moral principles, e.g. “Harming fish is morally wrong”. Personal values and preferences, such as the commitment to saving the fish, as well as third-party judgments are then derived from these general principles (see Figure 8).

Cultural learning → General moral principles → First-person values

Figure 8. The etiology of ‘principled moralization’.

But, alternately, development and socialization may shape our personal values and preferences directly (e.g., the aversion to cutting down trees, the commitment to the environment, and so on). Then, through projection, first-personal preferences and values shape our evaluations of others and our professed moral principles (see Figure 9). On this view, participants condemn David’s trade-off as a consequence of having a personal interest (a preference) in the preservation of fish. Either account seems capable of explaining the consonance of third-party evaluations and first-person protected values observed in Bartels’ (2008) study.

Cultural learning → First person values → General moral principles

Figure 9. The etiology of ‘projection’.

Some recent evidence provides support for the latter account (Lieberman & Lobel, 2012), showing that personal moral aversions concerning incest shape moral judgments of others’ incestuous behavior. Earlier research had shown that the duration of coresidence with siblings during childhood is one of the primary cues of relatedness (Lieberman, Tooby & Cosmides, 2007). So, Lieberman and Lobel (2012) set out to examine how coresidence duration might relate to one's personal aversion to incest and also to moral judgments about third-party incest. The researchers selected participants who had been raised in Israeli kibbutzim for two reasons: to confirm that this cue influences sexual aversions (1) towards peers who are not closely genetically related, and (2) even in the absence of social norms against sexual behavior towards co-reared peers. They found, as predicted, that duration of coresidence with an opposite sex peer predicted the strength of the aversion (i.e., disgust) toward sexual conduct with that same peer. It was also the case that participants' attitudes toward third-party sexual behavior between co-reared peers correlated with both total coresidence duration with opposite sex peers, and personal aversion towards sexual behavior with opposite sex peers.

Yet for this phenomenon to count as social projection, third-party judgments must *depend on* personal aversions. To test this claim using correlational data, it is common to employ what is known as a *mediation analysis* (Baron & Kenny, 1986). In a mediation model, you start with a known relationship (correlation) between two variables, and a direction of causation from IV (independent variable) to DV (dependent variable). For instance, we know that *age* is positively correlated with the amount of *savings in their bank account*. We can assume that as people get older (IV), this causes a probabilistic increase in their savings (DV), not *vice versa*. Having established this much, we ask whether the increase in savings can be attributed to a third, mediating variable (MV), *work years*, which is itself influenced by changes in the IV, age. The purpose of mediation analysis is to test the 'plausibility' of a causal model running from IV to MV to DV – and the extent to which MV mediates the relationship between IV and DV – by examining the patterns of shared variance between all three variables.

The researchers employed mediation to examine the causal relationships between coresidence duration, personal aversions to incest and third-party judgments. Since coresidence duration indexes a process taking place during childhood, one can assume that it precedes both the formation of sexual aversions and moral attitudes about third-party sexual behavior. Therefore, they posited coresidence duration as the independent variable in the model (i.e., the variable that causally influences other variables but is not causally influenced by them). This sets up the question about the relative order of own aversions and third-party moral attitudes in the causal chain. One possibility is that co-residence experience shapes own preferences and aversions, which in turn influence third party moral attitudes, as predicted by the projection view. Another possibility is that experience gives rise to the adoption of moral principles (indexed by third-party judgment), which then govern own aversions and preferences, as predicted by principled moralization.

In line with the projection model, controlling for own sexual aversions, the relationship between co-residence duration and moral attitudes disappeared. That is, first-person sexual aversions fully mediated the relationship between co-residence duration and moral disapproval of third-party peer sex, as predicted by the projection view. Meanwhile, counter to the principled moralization view, controlling for moral attitudes, the relationship between co-residence duration and personal aversions remained strong, indicating that moral principles do not (in this case) mediate the relationship between co-residence duration and personal sexual aversions. Therefore, the results of the comparison between mediation models favored the projection account, depicted in Figure 10.

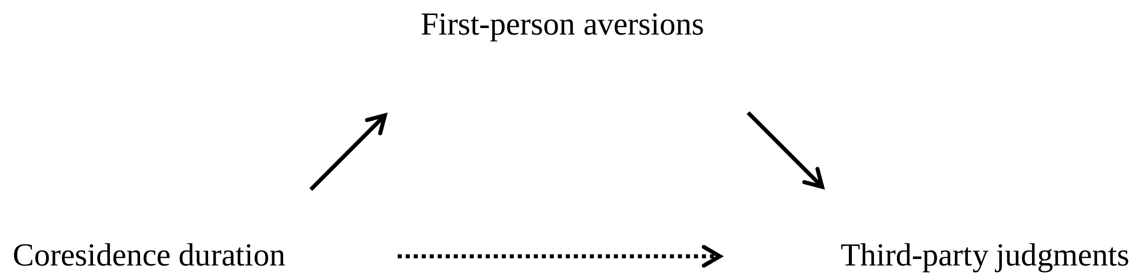


Figure 10. Mediation model: support for the projection view.

3.4. Simulation hypothesis: an alternative account?

The previous section advanced two broad empirical claims about the nature of spontaneous moral judgment:

> (1) the spontaneous processing of moral stimuli involves sensory and motor systems, and
>
> (2) moral judgments exhibit the typical pattern of projection, i.e., third-party moral attitudes are causally dependent on personal aversions and preferences.

Now we might ask: What sort of unconscious processes would recruit sensory and motor systems to visualize the action under evaluation, and cause third-party evaluations to be fashioned after own preferences and values? I suggest that processes of mental simulation are fit to accomplish just this. In the previous chapter, I introduced mental simulation as the imaginative ‘run-through’ of another’s behavior using shared neurocognitive systems. These shared systems are activated either endogenously or exogenously, such that consideration of others’ intentional actions, emotions, or sensations activates internal replicas of, among other things, the motor plans and affective states that lie behind the behavior (Gallese & Goldman, 1998; Gordon, 2004). I will refer to this account as the *simulation hypothesis*:

> Third-party moral judgments involve a (spontaneous) sensory and motor simulation, which elicits the evaluator’s personal affective response. This affective response shapes the moral evaluation of the third-party behavior.

Consider again the example of the footbridge problem. According to the above hypothesis, the presentation of the footbridge vignette yields – via mental simulation – an internal replica of the behavior, i.e., a first-person, imaginative run-through of the experience (in more or less detail). In turn, this process sets off an aversive, affective response which motivates the third-party judgment. This view would explain the involvement of motor systems, as suggested by Greene and colleagues' (2009) study, and of sensory systems as demonstrated by Amit and Greene (2011), and by Inbar and colleagues (2012).[10] In addition, the simulation hypothesis is equipped to explain why we find projection in moral evaluation of others (Lieberman & Lobel, 2007): the question whether another's behavior is right or wrong is answered by simulating own performance of the behavior, and morally evaluating another's action on the basis of the affective response that stems from the simulation.

Finally, I have argued that mental simulation may be the principal cognitive process in third-party moral judgment *and* that simulation (at least the simulation of emotionally-charged behavior) often results in an affective response. Therefore, a simulation-based account seems to favor a constitutive role of affect in moral judgment.

[10] It is worth also reviewing how this account would explain the correlational and experimental studies reviewed earlier in this chapter. First, consider studies that index a correlation between affect and third-party judgment. In the case of purity violations, for example, participants' own disgust sensitivity (arising through simulation) shapes condemnation of the third-party behavior. Similarly, the moral difference between *doing* a harmful action and *allowing* comparable harm to take place might arise from differences in the aversion to performing harmful actions versus performing harmful omissions (see 'omission bias' in Baron & Ritov, 1990; Ritov & Baron, 1995). Finally, on this view, the manipulations of affective state through hypnosis (Wheatley & Haidt, 2005) and watching comical videos (Valdesolo & DeSteno, 2008) influence the evaluator's actual affective state, interfering with the shared system's normal representation of the third-party agent's affective state. That is, these studies render the simulation mechanism less accurate at reflecting how the evaluator would feel when performing these third-party moral violations, thereby influencing the severity of their moral judgments.

That is, (if the account is right) we ought to conclude that moral judgments are constituted by affective responses to processes of third-party simulation.

3.4.1. The role of VMPFC in moral judgment

The proposed account faces a difficulty, prompted by the examination of moral judgment patterns in acquired sociopathy. Koenigs and colleagues (2007) tested the moral judgments of a group of patients with bilateral damage in the ventromedial prefrontal cortex (VMPFC) against a group of matched controls. They found that, compared to psychotypical participants in the control group, patients with VMPFC damage made more permissive judgments about harmful action in personal moral dilemmas. By contrast, the VMPFC patients made normal judgments about impersonal and non-moral dilemmas (see also Ciaramelli et al., 2007). VMPFC patients are known to have profound affective deficits, despite retained general intelligence. Remember that, on Greene's view, moral condemnation depends on affect-laden processes and rational processes support more permissive moral judgments about these cases. Therefore, Greene's theory straightforwardly predicts the result that is obtained. Meanwhile, on the account I have developed both kinds of judgments recruit affective processes, so it is not obvious how my account can explain these data.

A somatic marker view

A distinct perspective on the role of VMPFC, based in the work of Damasio and colleagues, suggests an alternative explanation for their moral judgments. According to the *somatic marker hypothesis*, psychotypical individuals have the ability to bring to bear affective cues in complex decision-making contexts, through a repository of unconscious associations between behavioral options and the outcomes that are associated through experience to each option (Damasio, 1996).

This perspective derives its strongest support from a series of experiments employing the Iowa Gambling Task. The IGT is a single player game where participants select cards from a set of decks seeking to maximize pay-off. Each card has a different pay-off but some decks have better average pay-off rates than others. Within a few rounds of the game, participants learn to select from the advantageous deck, before being able to consciously report which deck yields the highest rewards. Moreover, physiological evidence indicates that this process relies on anticipatory aversions that deter the selection from low-yield decks (Bechara et al., 1997). The VMPFC appears to play a crucial role in this process: Participants with lesions to the VMPFC fail (a) to learn to select from advantageous decks, and (b) to exhibit these same signs of anticipatory aversion to a selection from low-yield decks (Bechara et al., 1994). More recently, neuroscientific evidence points toward heightened activation in the VMPFC during the IGT (Li, Lu, D'Argembeau, Ng, & Bechara, 2010). Therefore, the VMPFC plausibly houses the somatic markers which unconsciously guide advantageous decision-making.

An analogous somatic marker account could be told about moral decision-making in trolley-type moral dilemmas. Somatic markers may also encode anticipatory aversions to immoral behavior, such as the violent action of pushing an innocent person. On this explanation, the personal (but not the impersonal) harmful action is associated through prior knowledge to ensuing victim harm, and therefore a fast and advantageous decision to condemn personal harm (but not impersonal harm) can be produced by psychotypical controls, owing to the influence of somatic markers.

Empathy and quantitative trade-offs

Still, it remains the case that VMPFC damage is associated with impaired affect, and moreover, impaired empathic abilities (Adolphs, 2002), and according to my proposal in Chapter 2 the endorsement of harm for the greater good recruits empathic abilities. So, why then would utilitarian judgment not also be impaired following

VMPFC damage (on my account)? Moreover, why would patients be *more* likely to endorse harmful actions for the greater good?

My account does indeed predict that the utilitarian concern with others' welfare should be subdued as a result of VMPFC damage. Reduced empathic concern should result in less motivation to actively "search" for moral outcomes and enact welfare-maximizing courses of action. This does appear to be true of their real-life behavior, as documented by the tendency toward antisocial personality disorder accompanying damage (Grafman et al., 1996) or cortical thinning (Narayan et al., 2007) of the VMPFC.

But then why does their reduced empathic concern go unnoticed in judgments about trolley-type dilemmas? Recall that in Chapter 2 we discussed the components of utilitarian judgment – probability and magnitude – and showed that it is determining the *magnitude* of moral outcomes which requires empathic abilities, but only (or primarily) when the relevant moral outcomes vary qualitatively. In addition, empathic concern may guide the prior "search" for morally relevant outcomes to consider in the welfare calculus. Yet, in the classic one vs. five moral dilemmas in the experimental literature, these morally relevant outcomes are (i) provided by the scenario description, and (ii) of a single kind. Even severely limited empathic concern might suffice to prefer the utilitarian course of action in these quantitative trade-offs. To see why, consider the following example:

I care deeply about my library. Therefore, if someone threatened to burn many of my books unless I shredded a single book from my library, I would adamantly comply. My CDs I don't like nearly as much; yet if someone posed the same dilemma, I reckon I would act the same (though not with the same drive, perhaps, with which I would try to save my library). I still prefer to save the larger number of my CDs, simply because I like CDs *at all*. This I think is roughly the case with VMPFC patients: they care less about human lives than do psychotypical utilitarians, but they still (in these particular circumstances) prefer welfare-maximizing outcomes. Their empathic deficit just means they won't pursue them with equal zeal.

In sum, VMPFC regulates a range of affective processes, including the influence of somatic markers in behavior regulation and the ability to empathize with others. Consequently, both the condemnation of harm *and* the promotion of the greater good are impaired following VMPFC damage. Guided by an experimental literature that has emphasized quantitative (*one-versus-five*) trade-offs, for which only minimal empathic concern is required, we have devised an incomplete picture of the role of VMPFC in moral judgment. Yet it is clear, outside lab settings, that VMPFC patients are not utilitarian in any broader sense. To the contrary, their empathic deficit renders them less concerned with the greater good, and less motivated to defend it.

3.5. Conclusion

A vast experimental literature demonstrates the involvement of affective processes in moral judgment. In this chapter, I differentiated three views about the psychological character of moral judgment, and argued that the empirical literature – and, in particular, the numerous studies employing experimental manipulations of affect – are incompatible with the epiphenomenal view, according to which affect merely results from the process of making moral judgments. Next, I addressed a related question concerning the cognitive processes that normally take place in the wake of third-party moral behavior, which yield this affective response. I characterized the prevailing account as one according to which the evaluator computes a moral intuition over different components of moral behavior such as intention, causation and so on. Instead, I argued that a growing body of literature points towards the involvement of a sensory and motor simulation of the behavior. A consequence of this is that we find a tendency to project one's own moral values and aversions in third-party moral judgment, as seen in experimental studies examining the moralization of protected values and the moralization of third-party incest among Kibbutz members. Therefore, I defend a view of the processing of moral stimuli according to which the evaluator spontaneously simulates the behavior under evaluation, and this process triggers (in some cases) aversive reactions that guide judgment and decision-making. If this account is correct (awaiting empirical confirmation), moral affect is the core mental state constituting spontaneous moral judgment.

Chapter 4. Testing the hypothesis of simulated aversion

4.1. Introduction

In the previous chapter I reviewed numerous studies demonstrating that moral judgments of purity and harm are widely influenced by affective processes. This is evident in the physiological and neural patterns accompanying moral judgment in psychologically typical populations (Greene et al., 2001; Moll et al., 2001, 2002a), in patients with blunted affect (Ciaramelli et al., 2007; Koenigs et al., 2007), and through the effects of transient manipulations of emotional state (Crockett et al., 2010; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005). I then argued that the data could be explained by appealing to processes of mental simulation, which shape our moral judgments about the behavior of others. However, direct empirical evidence for this phenomenon is limited. So, in this chapter, I aim to empirically test this suggestion, examining the precise content of this mental simulation.

A simple exercise reveals the surprising extent of this gap in our understanding. Imagine that you see a mother slap her child for dropping his ice cream on her foot. You judge the mother’s behavior to be immoral, and this depends in part on an affective response. What is its origin and nature? We distinguish between two possibilities, which are not mutually exclusive. According to the first, you focus on the mother and recoil at the thought of performing the act of hitting a child. According to the second, you focus on the child and cringe at the thought of his pain and humiliation. In other words, does your affective response arise from consideration of the act itself, or instead from consideration of its impact? Despite a wealth of evidence for the role of affect in the moral condemnation of harmful action, we have remarkably little basis on which to distinguish the contributions and roles of these two putative affective responses.

4.1.1. Two kinds of aversive affect

A dominant view, tracing back at least to Hume (1739), holds that the proscription of harmful behavior is rooted in a concern for the victim's distress (Hoffman, 1982, 2000; Pizarro, 2000) that emerges at an early age (Eisenberg-Berg, 1979; Smetana, 1985). The inhibition of violence depends largely on *empathic concern*, an affective response that stems from the apprehension of a victim's emotional state, and which is congruent with what the victim is feeling or is expected to feel as a consequence of the harmful act (Batson et al., 1993; Batson, 1994). As I discussed in Chapter 2, neuroscientific findings support the existence of empathy by showing similarity in the activation patterns during actual experience and observation in others of a range of emotional states, including pain (Morrison, Lloyd, di Pellegrino & Roberts, 2004). In laboratory investigations, these activations have been elicited by looking at facial (Carr, Iacoboni, Dubeau, Mazziotta & Lenzi, 2003) and bodily expressions (Jackson, Meltzoff & Decety, 2005), as well as by the mere imagination of another's experience (Singer et al., 2004; Jackson, Brunet, Meltzoff & Decety, 2006). Moreover, trait ratings of empathy are correlated with the strength of these activations (Lamm, Batson, & Decety, 2007; Singer et al., 2004). Critically, when this initial empathic response is activated, the agent's behavior can reflect increased concern for the victim's welfare. We call this the *outcome aversion* model: it posits an affective mechanism responsible for proscribing harmful behaviors based on the negative outcomes that they bring about, and is largely motivated by empathic concern for the victims (see also Miller, Hannikainen, & Cushman, 2013).

A conceptually distinct model, which we term the *action aversion* model, posits an affective mechanism that condemns actions intrinsically, rather than contingent upon the outcomes that they bring about. The clearest evidence favoring this model comes from cases of victimless crime. People condemn a host of actions, such as flag-burning and consensual incest, that do not directly cause victim distress (Haidt, Bjorklund, & Murphy, 2000; Haidt, Koller, & Dias, 1993). Consequently, the outcome aversion model is poorly suited to explain affective responses to violations of the purity domain. Instead, consideration of the action itself suffices to elicit the aversive affect that drives moralization.[11] Of course it is important to emphasize that the action aversion model concerns the *psychological mechanisms* that give rise to a particular judgment, not their developmental, historical or adaptive origins. It is surely the case, for instance, that sibling incest can lead to undesirable outcomes from a fitness perspective. Nevertheless, evidence suggests that our aversion to incest does not derive from the consideration of harmful outcomes at the level of psychological mechanism (Haidt et al., 2000; Lieberman & Lobel, 2012).

Violations in the harm domain present a more ambiguous case. Clearly, violently harming a person leads to bad outcomes (at least for the person harmed). Still, it is possible that the aversion to harmful action does not *require* consideration of the harmful outcome. Rather, it may depend in part on an intrinsic aversion to the action, including its sensory and motor properties; for instance, thrusting a knife at another person's belly.

Recent evidence supports the existence of action aversion in the harm domain (Cushman, Gray, Gaffey & Mendes, 2012). Specifically, individuals exhibited physiological signs of aversion to performing simulated harmful actions even when experimental conditions insured that the actions could not possibly cause harm. For example, participants shot an unloaded gun at an experimenter, or smashed a baby doll

[11] In positing the distinction between the agent's action and the outcomes to patients, our present framework neglects the contribution of a further potential factor in moral evaluation, i.e., the consideration of outcomes to the agent, e.g. remorse, shame. Indeed recent evidence suggests that these considerations influence the condemnation of purity violations relatively more than harm violations (Dungan, Chakroff & Young, in prep).

against a desk. The aversion elicited by performance of these actions was greater than it was witnessing an experimenter perform these same actions or performing metabolically matched control actions (e.g., squirting a spray bottle, hammering a block of wood). These results indicate an aversion to the performance of harmful action that is triggered even under conditions where it is known that no bad outcome will ensue, and only when the participant engages in *action* (not in cases matched for the imagined outcome).

Blair (1995, 2001, 2007) has developed a model of the neural and cognitive bases of psychopathy that provides a natural explanation for the action aversion hypothesis. According to Blair's *violence inhibition mechanism*, humans and other mammals possess an innate, unconditioned response to signs of victim distress that ceases ongoing behavior. Through a process of stimulus-reinforcement learning, actions that give rise to victim harm are associated with the negative reinforcement of the signs of victim distress. In this way, although outcome aversion may be the necessary developmental precursor of action aversion, consideration of the typically harmful action may independently suffice to elicit an aversion that deters harmful behavior.

4.1.2. Aversive affect in third-party moral judgment

Just as action- and outcome-aversion could operate as distinct affective bases for regulating one's own behavior, they may also contribute to processes of third-party moral judgment. For example, we may condemn someone else's choice to push the man off the footbridge on the basis of an aversion derived from considering the agent's action of pushing, or the victim's experience of being pushed off the footbridge. In this set of studies, we also address this question: To what extent do action and outcome aversion influence the moral evaluation of *others'* behavior? Our approach assumes that, to the extent that personal aversions are recruited in third-party evaluation, the mechanism likely involves perspective-taking. On the *agent focus* model, evaluative focus on the action, facilitated via simulation of the agent's perspective, triggers the basic affective response responsible for moral condemnation. In other words, agent focus uses one's own action aversion to inform third-party evaluation. In contrast, on the *victim focus* model, it is an evaluative focus on the outcome of the action (e.g. victim suffering), facilitated by simulation of the victim's perspective, that triggers the aversive affect that supports moral condemnation. Thus, victim focus uses one's own outcome aversion to inform third-party evaluation (see Table 4).

Table 4. *Agent and victim foci models.*

	Agent focus	**Victim focus**
Predominant affect:	Action aversion	Outcome aversion
Perspective-taking:	Agent perspective-taking	Victim perspective-taking

4.2. Are agent and victim simulation tractable in moral judgment?

We designed a questionnaire to assess individual differences in agent and victim foci during moral evaluation. The instrument is composed of two parts: the moral self-regulation section and the third-party evaluation section. With the moral self-regulation section, we employed an 18-item scale to assess the influence of action and outcome aversion on participants' moral decision-making. The items were divided into two subscales, *action focus* and *outcome focus*.

The *action focus* subscale contains nine items that stress the value of actions and decisions ("By and large, morality is about doing what feels right"), and assert the relevance of the agent's feelings ("At the end of the day, good moral decisions are those decisions you can live with") and attitudes ("Dignity is a big part of my morality, so there are certain things I could never do") for the decision-making process. In other words, they capture an approach to moral judgment in which the consideration of an action gives rise to a feeling that that action is right or wrong. The other nine items compose the *outcome focus* subscale, and stress the value of outcomes ("In order to be moral you have to pay close attention to the impact of your decisions"), and the relevance of welfare considerations ("Morality is about helping others and not harming them"), impartiality ("As far as morality goes my goal is to care about all people"), and allocentrism ("If only people cared more about each other, they would make better moral decisions") for moral decision-making. In contrast to the action focus items, every one of the outcome focus items made reference to the impact of actions upon others' welfare. In all the experiments that follow, the items were presented in an order randomized for each participant, and mixed across subscales, in order to preclude the impact of systematic order effects upon our analyses.

A salient dimension of divergence between the subscales is the influence of own feelings on moral decision-making: Action focus items often referenced the agent's feelings (e.g. "my own conscience", "dignity"), whereas outcome focus items did not. This reflects a basic distinction between action- and outcome-based approaches to moral judgment. In an action-based system, the moral status of an action is derived from the feeling elicited intrinsically by the action itself. By contrast, in an outcome-based system, the moral status of the action isderived from concern for the feelings and welfare of others (e.g., through empathy) and thus is contingent upon its consequences.

In the third-party evaluation section, participants reported their reliance on agent and patient perspective-taking as approaches to third-party judgment through a variety of measures linked to two short paragraphs. First participants viewed these paragraphs (randomized for order between participants), which contained short descriptions of agent perspective-taking (*Act*) and patient perspective-taking (*Impact*), and selected the paragraph that best described their own approach to the moral evaluation of third-party behavior:

> *Act*: I know what is right and wrong by listening to my own conscience. So, when judging another person's behavior, I put myself in their shoes and ask

> myself what I would have done. If I would have done the same, then what they did is morally right. But if I could not have done the same according to my conscience, then the behavior was morally wrong for someone else to do.
>
> *Impact:* I judge moral decisions by putting myself in the shoes of the people who are affected by those decisions. If performing an action hurts others, then I consider it morally wrong. If performing an action benefits others, I consider it morally acceptable. The primary purpose of morality is to help other people and not hurt them. Therefore, making a moral judgment is all about adopting the perspective of anybody who will be affected.

After endorsing one or the other paragraph, participants estimated the influence of agent perspective-taking (as defined by the Act paragraph), patient perspective-taking (as defined by the Impact paragraph) and any other approach they employ during third-party moral evaluation on independent scales. Additionally, participants indicated the relative influence that each approach had on their moral judgments using an 11-point bipolar scale, ranging from 1: ‘‘100% Act /0% Impact,’’ to 6: ‘‘50% Act /50% Impact,’’ to 11: ‘‘0% Act /100% Impact.’’ So, while the moral self-regulation scale was primarily diagnostic of an action- versus outcome-focused outlook, the measures on the third-party judgment page served to capture a participant’s explicit use of agent and victim perspective-taking in making moral judgments of the behavior of others.

Altogether, completion of the scale generated six indices for each participant. Two measures reflected action and outcome foci in moral self-regulation: (1) *action focus,* mean agreement with the nine action items on the scale, and (2) *outcome focus,* mean agreement with the nine outcome items on the scale. An additional four measures captured participants’ preference for agent or patient perspective-taking in third-party moral evaluation: (3) *agent vs. patient perspective-taking: endorsement,* whether participants endorsed the Act or Impact paragraph, (4) *agent perspective-taking: rating* and (5) *patient perspective-taking: rating,* participants’ ratings of the usefulness and importance of Act and Impact morality respectively, on a 7-point scale from 1: “not at all” to 7: “extremely”, and (6) *agent vs. patient perspective-taking: relative rating*.

4.2.1. Methods

Participants voluntarily logged on to the Moral Sense Test (moral.wjh.harvard.edu), a website hosted by the Psychology Department at Harvard University and maintained in collaboration with researchers at the Department of Cognitive, Linguistic & Psychological Sciences at Brown University, where visitors may take part in a range of experiments about moral psychology. The Moral Sense Test website has been used in previous studies of moral psychology (Cushman et al., 2006; Cushman, Knobe, & Sinnott-Armstrong, 2008; Hauser, Cushman, Young, Jin, & Mikhail, 2007). After a brief introduction to the nature and purpose of the study, participants completed our measures of agent and victim foci. At the end of experiment, participants optionally provided demographic information. 493 participants (276 females) completed the experiment.

4.2.2. Results

Action focus and outcome focus were moderately correlated, $r = .387$, $p < .0001$. An item analysis on action and outcome foci revealed that each subscale had satisfactory reliability, action focus $\alpha = .73$, outcome focus $\alpha = .73$. Every action item increased the reliability of the action focus subscale, and all but one outcome item ("If an action truly hurts nobody, then it probably isn't wrong.") increased the reliability of the outcome focus subscale.

An exploratory factor analysis initially retained seven factors with positive eigenvalues, of which only two factors presented eigenvalues > 1. A visual inspection of the scree plot also supported the extraction of two factors. The rotated factor analysis confirmed the scale's constitution of two factors corresponding to our *a priori* subscales: every item loaded > .3 on its corresponding factor and ≤ .3 on the other factor, except one outcome focus item (see Table 5). This outcome focus item (O5) articulated a typically consequentialist "no harm, no foul" view. We predicted that this item should be associated with outcome focus, where the welfare of victims and beneficiaries is seen as the basis of moral consideration, and yet it was an outlier on both the item and factor analyses. Consequently, in further administrations of this assessment of evaluative focus, we plan to exclude this item.

Table 5. *Rotated factor analysis: Factor loadings and uniqueness values.*

	Item	*F1*	*F2*	*Uniq.*
A1	By and large morality is about doing what feels right.	.532		.708
A2	When faced with a moral dilemma I usually listen to my own conscience.	.406		.811
A3	In a way, morality is like art. When you see something, you know how you feel about it.	.512		.719
A4	Dignity is a big part of my morality, so there are certain things I could never do.	.365		.839
A5	If only people listened to their inner voice, they would make better moral choices.	.487		.741
A6	At the end of the day, good moral decisions are those decisions you can live with.	.582		.633
A7	Something that feels repugnant for me to do is probably wrong for someone else to do.	.480		.754
A8	Certain ways of behaving are wrong no matter what the situation.	.356		.869
A9	If a behavior is morally right, it shouldn’t make me feel uncomfortable.	.531		.714
O1	I’m more than willing to make sacrifices for the better of others and the future.		.474	.765
O2	Morality is about helping others and not harming them.		.563	.642

O3	As far as morality goes, my goal is to care about people equally.		.459	.760
O4	The point of morality is to end suffering and promote happiness.		.535	.698
O5	If an action truly hurts nobody, then it probably isn’t wrong.			.954
O6	If only people cared more about each other, they would make better moral decisions.		.466	.755
O7	In order to live morally in your day-to-day, you have to constantly step out of your shoes.		.406	.834
O8	In order to be moral you have to pay close attention to the impact of your decisions.		.479	.768
O9	Considering the feelings of others is an important part of deciding what’s right.		.535	.688

Notes. F1: Factor 1 loadings; F2: Factor 2 loadings; Uniq.: uniqueness. Blanks represent factor loadings < .3.

Views on moral self-regulation predicted approaches to third-party moral evaluation. A logistic regression model on endorsement of agent vs. patient perspective-taking (1: agent; 0: patient) demonstrated unique effects of action focus, z = 8.02, p < .0001, and outcome focus, z = -11.01, p < .0001 (N = 1336, LR χ^2 = 156.80, r^2 = .089, p < .0001), indicating that action focus was associated with agent perspective-taking and outcome focus with patient perspective-taking. We confirmed these relations in multiple regression models on the independent ratings of agent and patient perspective-taking. There were unique effects of action focus, β = .366, p < .001, and outcome focus, β = -.138, p < .001, in the predicted directions on ratings of agent perspective-taking. Similarly, there were independent effects of action focus, β = -.130, p < .001, and outcome focus, β = .488, p < .001, on ratings of patient perspective-taking.

4.2.3. Discussion

In sum, responses to our measures of evaluative focus supported the dissociation between agent and patient foci. A rotated factor analysis revealed our two *a priori* factors in the moral self-regulation scale: one larger *outcome focus* factor on which all but one outcome item loaded and no action items loaded, and one smaller *action focus* factor on which all action items loaded, and no outcome items loaded. This finding indicates that participants' latent moral views contain these two distinct clusters of concerns. This distinction was further evinced by the relations between evaluative focus and perspective-taking in third-party evaluation: action focus was associated with adoption of the agent's perspective and outcome focus was associated with adoption of the patient's perspective. Together these analyses reveal (i) a distinction in participants' moral views between agent and patient foci, and (ii) individual differences in the reliance on each approach that are consistent across processes of self-regulation and third-party evaluation.

According to proponents of *dyadic completion*, the attributions of bad intentions to an agent and of suffering to a victim are essentially inseparable. So whenever a person perceives an agent acting immorally, they infer a patient who suffers; conversely, whenever they perceive a patient who suffers, they infer an agent acting immorally (Gray, Young & Waytz, 2012). This includes moral issues that do not have obvious agents, such as systemic injustice, and ones that don't have obvious victims, such as homosexual sex. This discounts a critical assumption of our experimental approach: i.e., that participants selectively simulation one or the other perspective. So, in order to rule out the possibility that individual differences on this dimension are an artifice of our stimuli, we asked a supplementary set of subjects (*N*=151) to report on the experience of completing our survey. An overwhelming majority (89%) of these participants found the choice between agent and patient perspective-taking either immediately obvious[12] (66%), or not immediately obvious but apparent after taking some time to reflect[13] (23%). This result suggests that the selective simulation of either agent or victim perspective is not an artifice of our stimuli.

In the following four experiments, we employ this scale to examine the roles of agent and patient foci in moral judgments of harm and purity violations. Our studies are arranged in pairs that focus on purity domain and harm domain violations, respectively. For the purity domain studies we have a strong *a priori* basis to predict that action aversion plays a larger role than outcome aversion because the violations are typically victimless; correspondingly, we also predict that agent perspective-taking plays a large role, with no meaningful role of victim perspective-taking. Consequently, within each pair of studies, the investigation of the purity domain serves partly as a method for validating the experimental paradigm used to assess action aversion and agent perspective-taking. For the harm domain studies we consider it an open question whether, and to what extent, agent and victim focus play a substantial role in moral judgment. We therefore leverage the experimental paradigms that we validated against the purity domain in order to assess the affective basis of moral judgment in the harm domain. In Experiments 1 and 2, we develop an individual differences measure to assess the use of action-focused and outcome-focused processes of moral judgment as well as agent versus victim perspective-taking in third-party evaluation. In Experiments 3 and 4, we manipulate agent versus victim perspective-taking to determine how these cognitive processes causally affect moral judgment.

[12] "As soon as I read both paragraphs, I knew which of the two paragraphs describes me better. So I found that answering this question was straightforward."

[13] "After reading both paragraphs, the answer was not immediately obvious. However, after thinking for a while, I realized that one paragraph describes me better than the other paragraph and I selected that one."

4.3. Experiment 1: Does the condemnation of purity rely on simulated aversion?

The first experiment assesses the role of agent and victim foci in moral judgments about violations of the purity domain, such as kissing a sibling in private, or eating one's dead pet dog (Haidt et al., 1993).

There is a strong *a priori* basis to predict that participants exhibiting a predominant action focus will tend to proscribe purity violations, whereas participants exhibiting a predominant outcome focus would tend to condone them, simply because most violations of the purity domain do not, or need not, involve a victim. For similar reasons, we predicted that perspective-taking would be associated with moral judgment, such that participants who tended to make moral judgments by adopting the agent's perspective would judge these purity violations more harshly than would those who tended to adopt the victim's perspective.

Our predictions add a twist to the well-established finding that condemnation of the purity domain is linked to the evaluator's disposition to experience disgust (Horberg et al., 2009; Inbar et al., 2009a; Rozin et al., 1999b). We propose that this relationship is due to individuals adopting an agent perspective, simulating performing the target behavior themselves, and using the elicited disgust response as a basis for moral judgment. Consequently, we predict that the established relationship between disgust sensitivity and condemnation of third-party purity violations will be moderated by perspective-taking; that is, disgust sensitivity should predict moral judgment significantly better among participants who tend to adopt the agent's perspective when making moral judgments.

4.3.1. Methods

Participants voluntarily logged on to the Moral Sense Test. After a brief introduction to the nature of the study, participants viewed a block of eight scenarios describing violations of the purity domain (see *Appendix A*), the measures of evaluative focus, and the Disgust Scale-Revised (DS-R; Haidt, McCauley & Rozin, 1994, modified by Olatunji et al. 2007), a reliable and widely used instrument for assessing individual differences in sensitivity to disgust that correlates with real-life behavioral responding to disgust-eliciting stimuli (Rozin et al., 1999a). The moral scenarios were presented in a pseudorandom order and briefly described a third-party agent performing actions that were designed to evoke core disgust (e.g., smearing feces on oneself), sexual disgust (e.g., French-kissing one's uncle at a family party), and animal-reminder disgust (e.g., getting plastic surgery that adds a two-inch tail to the end of one's spine). Participants rated the moral wrongness of these actions on a 7-point Likert scale from 1: "Not morally wrong at all" to 7: "Very morally wrong". The order of presentation of the scenario block and the measures of evaluative focus was counterbalanced, and participants always completed both sections before the DS-R scale. At the end of experiment, participants optionally provided demographic information.

437 participants (210 female) completed the experiment. Data were discarded from 52 participants (i) who completed the experiment in under 6 minutes (deemed the minimum time required for attentive participation), (ii) whose responses to our measures (action focus, outcome focus, ratings of agent and victim perspective-taking) and mean moral judgment deviated by over three standard deviations from the group mean, and (iii) who did not respond appropriately to the two catch items on the DS-R scale (e.g., by indicating disagreement with "I would rather eat a piece of fruit than a piece of paper.").[14]

4.3.2. Results

As hypothesized, action focus correlated with moral judgment, $r(388) = .28$, $p < .0001$, indicating that the more participants tended to moralize on the basis of actions

[14] All reaction time filters in this chapter were established by a collaborator on these studies. In order to establish a minimum time, he read through the each question at a fast pace and completed each question shortly after reading the prompt.

the more they condemned third-party violations of the purity domain. This correlation was not significantly influenced by the order of presentation of the two measures, *p* > .9, and in a one-way ANOVA (*condition*: EF before, EF after) on moral judgment, there were no effects of condition, *p* > .5. By contrast, no relationship was found between moral judgment and outcome focus, *p* > .5 (see Figure 11).

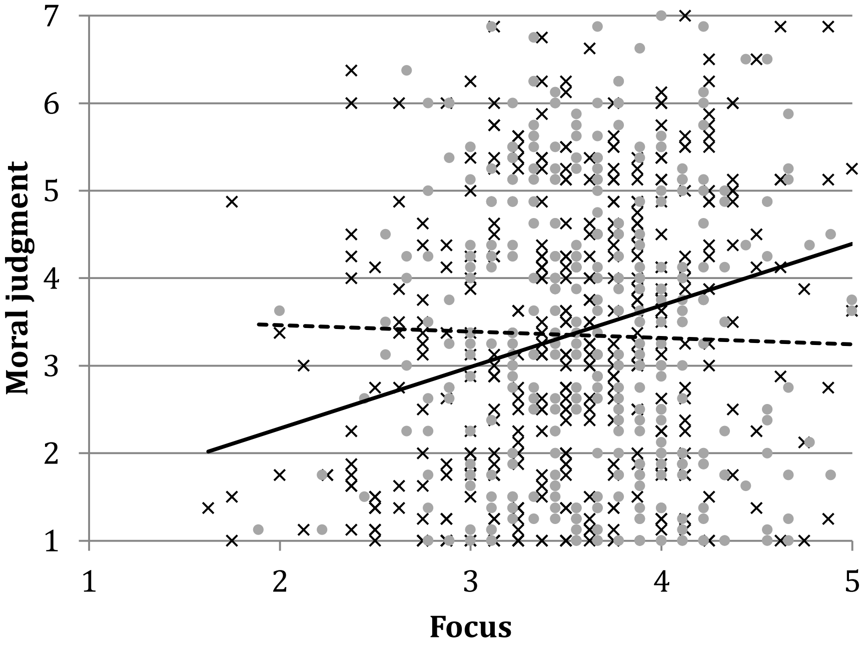


Figure 11. Moral judgment by action (x; solid trend line) and outcome (o; dotted trend line) foci.

We also found the predicted difference in moral judgment between participants who endorsed agent versus victim perspective-taking as approaches to third-party moral judgment. Participants who reported adoption of the agent's perspective made harsher moral judgments (*M* = 3.63, *SD* = 1.64) than did participants who reported adoption of the victim's perspective (*M* = 3.10, *SD* = 1.52), *t*(386) = 3.29, *p* < .002. Similarly we found a positive correlation between the rating of agent perspective-taking and moral judgment, *r*(388) = .16, *p* < .002. This correlation was, however, dependent on the order of presentation: EF after *r*(188) = .32, *p* < .0001; EF before *r*(200) = .01, *p* = .9. A similar relationship was observed between the relative rating of agent versus victim perspective-taking and moral judgment: EF after $r(188) = -.23$, $p < .001$; EF before $r(200) = -.09$, $p = .2$. Ratings of victim perspective-taking did not correlate with moral judgment, $p > .3$.

In accordance with previous studies, we found also that participants' disgust sensitivity correlated with their moral judgments, $r(388) = .42$, $p < .0001$. We hypothesized that this relationship is the result of participants simulating the agent's perspective. According to our description of agent perspective-taking in third-party evaluation, third-party behavior is condemned when simulation of the action in question elicits an aversive response (in this case presumably a disgust response). Critically, however, the disgust response should only serve as a basis for the condemnation of third-party behavior to the extent that participants adopt the agent's perspective—that is, to the extent that they take their own feeling of disgust as relevant to judging another's behavior. In order to test this hypothesis, we examined the pairwise correlations between moral judgment and disgust sensitivity, comparing participants who reported adopting the agent's perspective to participants who reported adopting the victim's perspective. Indeed a Fisher's r-z test revealed that the correlation between disgust sensitivity and moral judgment was significantly stronger among agent perspective-takers, than among victim perspective-takers, $z = 2.71$, $p < .007$ (agent $r(165) = .54$, $p < .0001$, victim $r(223) = .31$, $p < .0001$).[15] This result supports our hypothesis: proneness to feeling disgust in response to aversive stimuli accounted for 29% of the variance in moral judgment among participants who reported adopting the agent's perspective but only 10% of the variance among participants who reported generally endorsing the victim's perspective. We confirmed this difference by entering disgust sensitivity, endorsement and their interaction into a multiple regression model predicting moral

[15] The difference in these correlations is evidently not affected by order: *EF first* – agent $r(92) = .58$, $p < .0001$; victim $r(108) = .26$, $p < .007$; *EF after* – agent $r(73) = .51$, $p < .0001$; victim $r(115) = .33$, $p < .0005$.

judgment, $F(3,388) = 33.70$, $r^2 = .208$, $p < .0001$. We found main effects of disgust sensitivity, $\beta = .40$, $p < .001$, and endorsement (agent vs. victim perspective), $\beta = .13$, $p < .003$, and also critically the predicted interaction, $\beta = .13$, $p < .004$ (see Figure 12), indicating that the effect of disgust sensitivity on moral judgment was moderated by dispositional perspective-taking in third-party moral judgment.

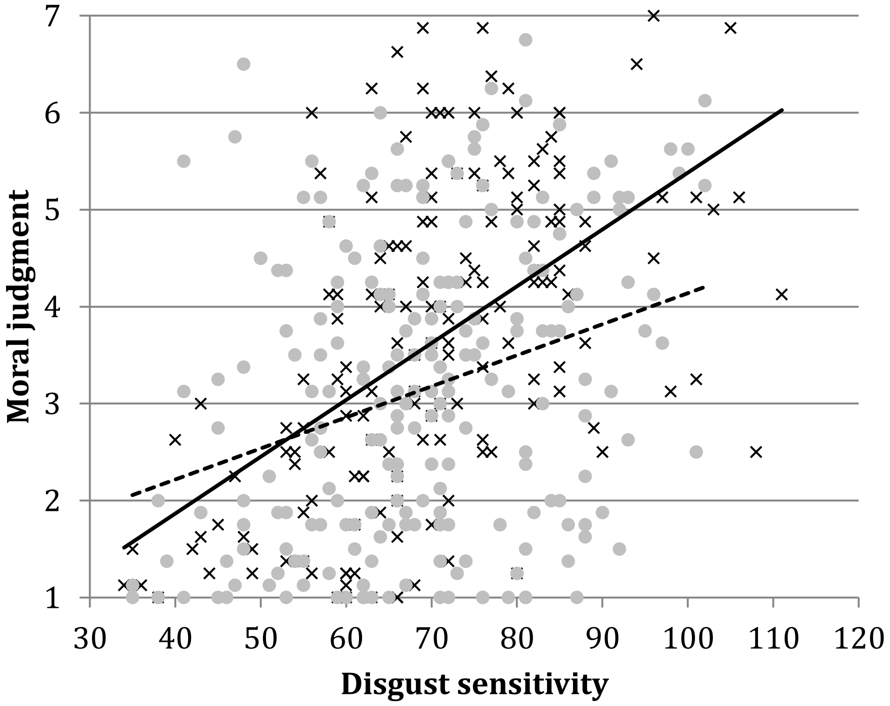


Figure 12. Disgust sensitivity and moral judgment by perspective-taking (agent: x, solid line; victim: o, dotted line).

We replicated established relationships between political orientation, religiosity and moral judgment of purity violations: moral judgment correlated with religiosity, $r(388) = .48$, $p < .0001$, and political orientation, $r(374) = .39$, $p < .0001$, indicating that conservative and religious participants tended to condemn purity violations relatively more than did liberal and non-religious participants. In addition, we discovered relationships between religiosity and political orientation and evaluative foci, on both measures of self-regulation and third-party judgment. Action focus correlated with political orientation, $r(374) = .23$, $p < .0001$, and religiosity, $r(388) = .22$, $p < .0001$, indicating that conservative and religious participants tended to have a greater action focus. Outcome focus did not correlate with religiosity, *p* > .7, or political orientation, *r*(373) = -.07, *p* = .15, though there was a small trend towards greater outcome focus among liberals.

We found corresponding relationships also for perspective-taking in third-party moral judgment. Participants who endorsed victim perspective-taking were significantly more politically liberal and less religious (pol. *M* = 3.22, *SD* = 1.45; rel. *M* = 2.96, *SD* = 2.03) than participants who endorsed agent perspective-taking (pol. *M* = 3.76, *SD* = 1.59; rel. *M* = 3.55, *SD* = 2.02), pol. *t*(372) = 3.40, *p* < .0007, rel. *t*(386) = 2.81, *p* < .005. These relationships were reflected also on the relative ratings of agent vs. victim perspective-taking, pol. *r*(374) = -.21, *p* < .0001; rel. *r*(388) = -.13, *p* < .01. Similarly, ratings of agent perspective-taking correlated with political orientation, *r*(374) = .17, *p* < .0007, and religiosity, *r*(388) = .20, *p* < .0001, such that conservative and religious participants rated agent perspective-taking more favorably than did liberal and non-religious participants. In addition, victim perspective-taking was associated with a liberal political orientation, *r*(374) = -.13, *p* < .01.

4.4. Experiment 2: Does the condemnation of harm rely on simulated aversion?

In this experiment we employed our measures of evaluative focus in order to explore the relationship of agent and victim focus to moral judgments of hypothetical dilemmas that involve a tradeoff among lives, such as the *trolley problem* (Foot, 1967; Thomson, 1985). Our dilemmas described a person (the “agent”) who brought about direct harm to one person (the “proximal victim”) in order to save a greater number of other people (the “distal victims”). We contrasted two types of scenarios that have been extensively investigated in past research (Cushman et al., 2006; Greene et al., 2001 2009; Mikhail, 2000; Petrinovich et al., 1993): *personal* cases, in which the agent brought about the harm as a means to saving the five and by applying forceful contact (e.g., pushing someone off the footbridge in order to stop the train, thereby saving the five), and *impersonal* cases, in which the harm brought about by the agent involved neither forceful contact nor was a means to saving the five (e.g., turning a train onto a sidetrack, where the victim is standing and dies as a side-effect of saving the five). As I discussed earlier, people typically judge that killing the proximal victim is morally wrong in the personal cases, but morally permissible in the impersonal cases.

Experiment 2 asks whether, in the personal cases, moral condemnation is associated to agent or victim focus. On the victim focus view, the critical difference between personal and impersonal cases is in empathic concern towards the proximal victim, and this difference is supported by a simulation of the victim's perspective. Several considerations support this view. First, the proximal victim seems relatively more 'prominent' in the personal cases: when we contemplate the footbridge case, we imagine the victim being forcefully harmed "up-close", and this may elicit greater empathic concern. Second, when we perceive intention in the perpetrator's attack, the attack seems more painful than if it had been inflicted accidentally (Gray & Wegner, 2008). This effect would result in the perception of greater pain, and likely more empathic concern for the victim in the personal cases, where the agent is described as harming the victim *in order to* save the five, than in the impersonal cases, where the victim is harmed only *as a side effect* of saving the five. Third, populations known to have deficits in empathic concern, as indexed by blunted physiological (Blair, Jones, Clark, & Smith, 1997; Damasio, Tranel & Damasio, 1990) and neural (Deeley et al., 2006) reactions to the perception of sad and fearful faces, demonstrate abnormally high rates of endorsement of personal harm (Bartels & Pizarro, 2011; Koenigs et al., 2007, 2012). Thus, prior work suggests that outcome focus, facilitated by victim perspective-taking, may contribute to the enhanced condemnation of personal harm.

On the agent focus view, an evaluation focused on the agent's action gives rise to the difference in moral judgment between these types of dilemmas. Specifically, moral condemnation depends on greater aversion towards performing the agent's action in the personal version than the impersonal version, and this process is supported by a simulation of the agent's perspective. A recent experiment examining the features that drive the condemnation of personal harm supports this view (Greene et al., 2009). Specifically, condemnation of the welfare trade-off is triggered not by spatial proximity or physical contact between the agent and the victim, but by the application of the agent's muscular force to the victim. Moreover, this effect is heightened when the agent acted intentionally, such that participants condemned the welfare trade-off most when the agent applied muscular force to bring about harm *as a means to* saving the five. These findings point towards "a system of moral judgment that operates over an integrated representation of goals and personal force—representations such as 'goal-within-the-reach-of-muscle-force.'" (Greene et al., 2009, p. 370), suggesting that moral judgment is influenced by an evaluation of the *agent's* motor behavior and goal, rather than the victim's perceived suffering. Furthermore, evidence that people show an aversion to performing simulated harmful actions even when they give rise to no actual harm (Cushman et al., 2012) indicates a potential basis for the action-focused affective processes engaged in the moral judgment of dilemmas involving harm.

In sum, when we contemplate the trolley problem and judge that it is wrong to push the man but ok to flip the switch, are we feeling greater compassion for the man on the bridge than we feel for the victim on the sidetrack, or are we feeling greater aversion to the thought of pushing a person than we feel towards pulling a switch? In this experiment, we take a first step in answering this question by examining whether individual differences in agent or victim focus are selectively associated with differences in the condemnation of personal harm.

4.4.1. Methods

Participants voluntarily logged on to the Moral Sense Test website. In a 2 (*Personal* vs. *Impersonal* moral dilemmas) x 2 (EF *Before* vs. *After*) between-subjects design, participants viewed six moral dilemmas (see *Appendix B*) and completed the indices of evaluative focus. Moral dilemmas described an agent who brought about harm to a proximal victim and saved a greater number of people. Participants then rated the moral wrongness of the agent's behavior on a 7-point Likert scale from 1: "Not morally wrong at all" to 7: "Very morally wrong". Dilemmas were presented in a pseudorandom order and the order of presentation of the dilemmas block and the measures of evaluative focus was counterbalanced. Finally, participants optionally provided basic demographic information.

4.4.2. Results

425 participants (223 female) completed the experiment. Data were discarded from 38 participants who (i) completed the experiment in under 4 minutes (deemed the minimum completion time), and (ii) whose responses to our measures (action focus, outcome focus, ratings of agent and victim perspective-taking) and mean moral judgment deviated by over three standard deviations from the group mean.

In a two-way 2 (*condition*: Personal, Impersonal) × 2 (*order*: EF before, EF after) ANOVA on moral judgment, there was a main effect of condition, $F(1,391) = 125.1$, $p < .0001$, and no effect of order of presentation or interaction with condition, *p*s > .3. The main effect of condition indicated that participants rated the welfare trade-off as morally worse in the personal condition ($n = 204$, $M = 4.10$, $SD = 1.18$) than in the impersonal condition ($n = 187$, $M = 2.80$, $SD = 1.05$), $t(389) = -11.45$, $p < .0001$.

Turning then to the main analysis of interest, we found that action focus correlated with deontological moral judgment in the personal condition, $r(204) = .25$, $p < .0004$, whereas outcome focus did not, $r(204) = .08$, $p = .3$. On impersonal dilemmas, the correlation between action focus and moral judgment approached significance, *r*(187) = .12, *p* = . 10, and there was no relationship between moral judgment and outcome focus, *r*(187) = .06, *p* = .4 (see Fig. 3). A multiple regression model predicting moral judgment by dilemma type, action focus, and the dilemma type x action focus interaction revealed main effects of action focus, *β* = .17, *p* < .001 and dilemma type, *β* = .48, *p* < .001, but the interaction term did not reach significance, *β* = .07, *p* = .12. In addition, the correlation between action focus and moral judgment was stronger when participants completed the scale prior to the moral dilemmas: EF first *r*(92) = .34, *p* < .001; EF after *r*(112) = .17, *p* < .07.

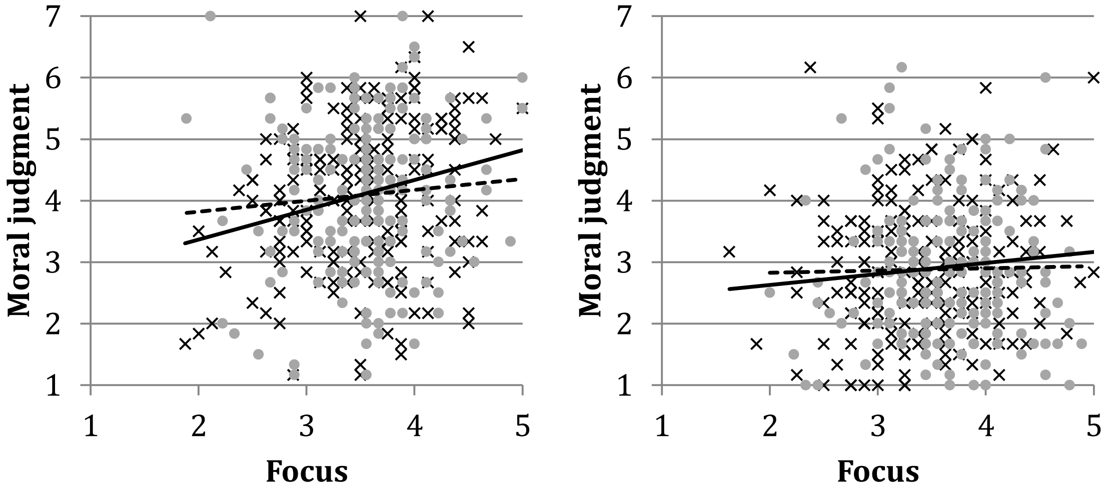


Figure 13. Moral judgment by action (x, solid trend line) and outcome foci (o, dotted trend line): Personal moral dilemmas (left), impersonal moral dilemmas (right).

Our presumption that action focus in third-party moral judgment should be accomplished by agent perspective-taking was supported by evidence of a correlation between these two measures. Consequently, we predicted that agent perspective-taking should be associated with deontological moral judgment on personal harm dilemmas. However, we found no correlation between moral judgment and either endorsement or ratings of agent and victim perspective-taking, all *p*s > .7.

Taken together, these results provide partial support for the action model in the harm domain. Action focus in moral self-regulation correlated with the tendency to condemn harmful action in personal dilemma contexts, but not impersonal dilemma contexts. A related prediction concerning self-reported perspective-taking in third-party moral judgment was not supported.

We replicated relationships between religiosity and political orientation and responses to the assessment of evaluative focus found in Experiment 1. Political orientation correlated negatively with outcome focus, *r*(369) = -.22, *p* < .0001, and positively with action focus selectively in the *EF first* condition, *r*(177) = .22, *p* < .004. Similarly, political orientation correlated negatively with ratings of victim perspective-taking, *r*(369) = -.17, *p* < .001, and positively with ratings of agent perspective-taking, *r*(369) = .18, *p* < .0005. Religiosity correlated with action focus, *r*(390) = .29, *p* < .0001, and agent perspective-taking, *r*(390) = .25, *p* < .0001, but did not correlate with outcome focus or victim perspective-taking, *p*s > .8. Lastly, on our measure of endorsement of agent vs. victim perspective-taking in third-party moral judgment we found the same relationship: participants who endorsed victim perspective-taking were significantly more liberal (*n* = 199, *M* = 3.01, *SD* = 1.46) than were participants who endorsed agent perspective-taking (*n* = 170, *M* = 3.75, *SD* = 1.69), *t*(367) = 4.56, *p* < .0001. The corresponding difference in religiosity trended in the same direction, but non-significantly, *t*(388) = 1.53, *p* < .13. In sum, we found consistent patterns of relationship on both measures of self-regulation and third-party judgment; namely, conservatives and religious participants tended to be agent-focused while liberals tended to be victim-focused.

4.5. Discussion

Two experiments provide preliminary support for the role of agent focus in the proscription of purity violations and of personal harm in moral dilemma contexts. We found that a tendency to focus on actions, but not outcomes, in moral self-regulation correlated with moral judgment of third-party purity violations and personal moral dilemmas, such as the footbridge dilemma. Participants who exhibited a greater focus on actions tended to judge these violations more severely than did participants who exhibited lesser action focus.

We found also that action and outcome foci in moral self-regulation covaried with adoption of the agent's and victim's perspective respectively in third-party moral judgment. Therefore, we predicted that self-reported agent perspective-taking in third-party evaluation would be associated with moral judgment. However, we found limited support for this hypothesis. Participants who endorsed agent perspective-taking as an approach to third-party judgment condemned violations of purity more than did participants who endorsed victim perspective-taking, but only when the perspective-taking questions followed the judgment task. We also found that agent perspective-taking moderated the effect of disgust sensitivity on condemnation of purity violations. A relationship between self-reported perspective-taking and moral judgment was not observed, however, in the harm domain.

In Chapter 3 I differentiated between the categorization view and the simulation view. Notice that either explanation can account for our pattern of results: i.e., that moral judgment was associated with our measure of action focus, but not with our measure of agent perspective-taking. On the *categorization/computation* view, certain actions "feel" categorically wrong, whether carried out by the self or others. Certain moral principles determine that personal harm is wrong, and then these principles cause our preferences and affective responses. Consequently, action focus would predict moral judgments of personal harm, but without employing processes of simulation or perspective-taking. By contrast, on the simulation view, certain actions feel wrong to the evaluator and, through a mental simulation of the agent's perspective, are evaluated as wrong for others.

So the simulation view would have predicted a relationship between our measures of perspective-taking and moral judgment. However, as prior research has shown, it is often the case in social cognition that cognitive processes are unconscious, and inaccessible by introspection (Cushman et al., 2006; Hauser et al., 2007; Inbar et al., 2009b). In our present case, if perspective-taking in third-party moral judgment is relatively unconscious, this would explain the failure of our self-reported measure of perspective-taking. So in Experiments 3 and 4 we pursue a further test of these distinct hypotheses.

We found no relationship between moral judgment and the tendency to focus on outcomes or to adopt the victim's perspective in third-party moral evaluation. One explanation for this might be the presence of outcome considerations motivating both attitudes of condemnation and of exculpation. That is, one could in principle focus on the proximal victim and heavily condemn the welfare trade-off, or instead empathize with the distant five and endorse it. Consequently, across trials and individuals, we might have found no correlation between moral judgment and outcome focus, even though outcome focus and victim perspective-taking may promote *both* judgments. So in Experiment 2b we take a different approach and examine the effect of *proximal* victim perspective-taking, rather than of general, dispositional outcome focus and victim perspective-taking.

Although not an *a priori* concern of our study, we found consistent relationships between both political orientation and religiosity, and agent versus victim foci in moral judgment. Namely, religiosity and political conservatism correlated with a tendency to focus on actions and to simulate the agent's perspective while liberalism was associated with a focus on outcome considerations and with simulating the victim's perspective during moral evaluation. This finding dovetails with evidence that liberals principally moralize domains that directly implicate harm to others – i.e., the harm and fairness domains – whereas conservatives tend to regulate a wider range of behaviors, many of which do not obviously bring about harmful outcomes (Graham, Haidt, & Nosek, 2009). This result will be the focus of the following Chapter 5.

4.6. Perspective induction via narration

In Experiments 3 and 4 we target the role of perspective-taking in third-party moral evaluation. We did not find strong evidence for this relationship in our previous experiments, and the results produced can be explained by either categorization or simulation views. I argued though that this may be due to the limitations of a self-report paradigm, since mental simulation in moral judgment could be unconscious. Consequently, we adopt a direct manipulation of perspective-taking in the following experiments.

Our methodological approach in Experiments 3 and 4 depends on the use of narrative focus to manipulate participants' perspective on moral stimuli. Theorists in the psychology of narrative have argued that, in order to process narrated events, readers assume the perspective of a character (Black, Turner, & Bower, 1979; Özyürek & Trabasso, 1997; Rall & Harris, 2000) and mentally represent the emotional states (Gernsbacher, Goldsmith, & Robertson, 1992), beliefs and goals (Wegner & Giuliano, 1983) and motor plans (Mar, 2004) of that character. In this way, the reader may experience affect congruent with the character's situation and comparable to the emotional experiences encountered in the real-world (László & Cupchik, 1995; Oatley, 1999). Neuroscientific evidence of *shared systems* provides support for these claims. Observation of third-party action activates neural systems that subserve the observer's performance of that action (Gallese, Keysers & Rizzolatti, 2004; Rizzolatti & Craighero, 2004), and it recently been demonstrated that reading can activate shared systems also (Aziz-Zadeh, Wilson, Rizzolatti & Iacoboni, 2006). Importantly, shared systems have been shown to activate during both observation of motor behavior (Cochin, Barthélémy, Roux & Martineau, 1999; Rizzolatti et al., 1996) and the recognition of affective states (Carr et al., 2003; Wicker et al., 2003).

4.7. Experiment 3: Does perspective-taking influence condemnation of purity violations?

In Experiment 3, we employed a perspective manipulation in order to examine whether violations of the purity domain are judged differently when narrated from the agent's point of view versus the point of view of a passive bystander, while matching descriptions of the action plan for length and detail. This manipulation should affect the extent to which participants simulate the agent's behavior and consequently the salience of the resulting action aversion. Therefore, if a mental simulation of the agent's action drives condemnation of violations of the purity domain, the behavior should be perceived as morally worse from the agent's perspective than from a bystander's perspective.

4.7.1. Methods

Participants voluntarily logged on to the Moral Sense Test website, were randomly assigned to one of two conditions (*Agent* or *Bystander*) and viewed a random selection of four scenarios drawn from a set of eight total scenarios (see *Appendix C*). The vignettes described violations of the purity domain, beginning with an introduction that served to establish the participants' tacit identification either as the agent or a bystander in the scenario. Consider, for example, the following scenario involving the artists Julie (agent) and Meg (bystander; italicized content varies by condition):

> *Julie/Meg* is an artist who lives in the heart of Boston with her husband and two children. One spring morning, *Julie/Meg* wakes up early to meet a friend and fellow artist at a nearby café for breakfast. *Julie/Meg* decides to buy a cup of coffee and a blueberry muffin before she sits down with her friend. They discuss collaborating on a new project and a proposed budget for the project. On the way to the art studio, *Julie/Meg's friend* notices city staff putting up decorations

> to celebrate the upcoming religious holiday. When they arrive at the art studio, *Julie's friend/Meg* begins to brew a pot of tea.

After the perspective-inducing introduction, the action plan was presented:

> As *her friend/she* is making the tea, *Julie/Meg's friend* prepares her workstation and says she wants to draw a portrait of the religious icon she saw on the way to the studio. However, she would depict him nude and make the portrait rather comical with the religious leader's genitalia appearing grossly enlarged.
>
> How morally wrong would it be for *Julie/Meg's friend* to make this portrait?

Participants then made moral judgments by selecting their response on a 7-point Likert scale from 1: "Not morally wrong at all", to 4: "Somewhat morally wrong", to 7: "Very morally wrong". After the block of purity violations, participants completed the assessment of evaluative focus and provided optional demographic information.

4.7.2. Results

660 participants (303 female) completed the experiment. Data were discarded from 24 participants who (i) completed the experiment in under 4 minutes (deemed the minimum completion time), and (ii) whose responses to our measures deviated by over three standard deviations from the group mean. After applying this filter, 321 participants remained in the Agent condition, and 315 in the Bystander condition.

Because of the design of Experiments 3 and 4, we conducted all our analysis replacing random effects tests with multilevel tests, where we entered scenario context and subject as dummy variables. A multilevel mixed-effects linear regression on moral judgment with condition (1: Agent, 0: Bystander) as independent variable, and scenario context as dummy variable revealed a significant effect of condition, $z = 2.16$, $p = .031$, indicating that participants tended to condemn purity violations more in the Agent condition than in the Bystander condition (see Figure 14). Looking at individual scenarios, we found the predicted difference in moral judgment on seven of the eight scenarios (see Table 6).

Table 6. *Moral judgment by scenario and condition.*

Scenario	*Agent*			*Bystander*		
	N	Mean	SD	N	Mean	SD
Smearing feces on self	157	3.78	2.45	146	3.19	2.34
Drawing a naked portrait of religious leader	152	3.73	2.13	146	3.36	2.18
Having sex with a chicken carcass	147	4.27	2.40	154	4.51	2.50
French-kissing a family member	159	4.89	2.11	159	4.64	2.25
Signing a note to sell her soul	158	3.56	2.42	158	3.42	2.43
Urinating on someone at their request	156	3.26	2.26	156	2.97	2.24
Eating a sandwich containing own pubic hair	161	3.47	2.43	159	3.32	2.46
Adding a two-inch tail to one's spine	154	3.12	2.26	159	3.08	2.26

Collapsing across conditions, we replicated the relationships observed in Experiment 1 between evaluative focus and moral judgment. First, action focus was associated with greater condemnation, z = 10.31, p < .001, whereas outcome focus was not, p > .8. We observed the corresponding relationships also with perspective-taking: participants who endorsed agent perspective-taking made harsher moral judgments than did participants who endorsed victim perspective-taking, z = 4.82, p < .001. Moreover, the higher participants rated agent perspective-taking, the harsher their moral judgments tended to be, agent z = 9.04, p < .001, relative agent vs. victim z = -5.79, p < .001. We uncovered a relationship in the opposite direction with victim perspective-taking: the more participants tended to adopt the victim's perspective, the less they condemned purity violations, z = -2.05, p < .04.

We replicated also the relationships between political orientation, religiosity and evaluative foci measures found in Experiments 1 and 2: Participants who endorsed victim perspective-taking were significantly more politically liberal and less religious (*n* = 404; pol. *M* = 3.04, *SD* = 1.43; rel. *M* = 2.81, *SD* = 1.94) than participants who endorsed agent perspective-taking (*n* = 232; pol. *M* = 3.51, *SD* = 1.50; rel. *M* = 3.27, *SD* = 2.06), pol. *t*(603) = 3.77, *p* < .0002, rel. *t*(631) = 2.85, *p* < .005 (see Table 7 for correlations).

Table 7. *Political orientation, religiosity and evaluative focus: correlations.*

	Focus		*Perspective-taking*		
	Action	Outcome	Agent	Victim	Relative
Political conservatism	.21 ***	-.16 ***	.11 *	-.11 **	-.13 **
Religiosity	.28 ***	.04	.20 ***	-.04	-.13 **

*: *p* < .05; **: *p* < .005; ***: *p* < .0005

4.8. Experiment 4: Does perspective-taking influence condemnation of harm?

Experiment 3 yielded support for the role of agent perspective-taking in the condemnation of violations of the purity domain. In this experiment, we employ the same perspective manipulation to examine the effect of perspective-taking on moral judgments of personal harm dilemmas. If agent perspective-taking drives the condemnation of personal harm, we should expect moral judgments to be more deontological from the agent’s perspective than from a bystander’s perspective.

The design of Experiment 4 allows us to address an additional important issue not resolved by Experiment 2. In Experiment 2, we found no relationship between moral judgment and the tendency to adopt the victim’s perspective in third-party moral evaluation. We proposed that this may have been due to the presence of multiple potential victims in a typical moral dilemma. So, in this experiment we isolate the effects of *proximal* victim perspective-taking in order to ask whether, and to what extent, empathizing with the proximal victim influences condemnation of the welfare trade-off in personal dilemma contexts. If considerations of victim pain and suffering drive condemnation of personal harm, then participants who read a scenario narrated from the proximal victim's point of view should condemn harm directed toward that victim particularly strongly.

In summary, Experiment 5 assesses the roles of agent and victim perspective-taking in the condemnation of personal harm. In simple terms, when we ponder the footbridge dilemma, what is the basis of our condemnation? Are we imagining what it would be like to be run over a train, or instead what it would be like to push a person off a footbridge? [16]

4.8.1. Methods

Participants voluntarily logged on to the Moral Sense Test website and were randomly assigned to one of three conditions: *Agent*, *Victim* or *Bystander*. In an incomplete repeated-measures design, participants viewed four personal moral dilemmas drawn from a total set of six (see *Appendix D*). In each condition, the vignettes were furnished with extended introductions that served to induce the perspective of the agent, victim, or bystander. After the perspective-manipulation, the dilemma was presented and participants were asked to make moral judgments (e.g., "For Brooke to throw the old man overboard would be…") by selecting a response on a 7-point Likert scale from 1: "Not morally wrong at all", to 4: "Somewhat morally wrong", to 7: "Very morally wrong". After the moral dilemmas section, participants completed an adapted version of the "Linguistic Implications Form", a pronoun

[16] This experiment included a pronoun completion task (Wegner & Giuliano, 1980) in order to measure self-focused attention. Participants completed this task immediately following the moral judgment task. However, this task did not generate noteworthy results, so it is excluded from further discussion.

completion task developed by Wegner and Giuliano (1980) measuring self-focused attention. Lastly, participants completed the assessment of evaluative focus and provided optional demographic information.

4.8.2. Results

1002 participants (467 female) completed the experiment. Data were discarded from 24 participants who (i) completed the experiment in under 6 minutes (deemed the minimum completion time), and (ii) whose responses deviated by over three standard deviations from the group mean. After applying this filter, we had 326 participants in the Agent condition, 317 in the Victim condition, and 335 in the Bystander condition, roughly matching the sample sizes employed in Experiment 3.

A multilevel mixed-effects linear regression on moral judgment with Agent condition (1: Agent, 0: Victim, Bystander) and Victim condition (1: Victim, 0: Agent, Bystander) as the independent variables (and scenario context and subject as random-effects) revealed a significant effect of Agent condition, $z = 2.98$, $p = .003$, but no effect of Victim condition, $z = 1.20$, $p = .23$. This result indicates that participants in the Agent condition condemned personal harm more harshly than did participants in the Bystander condition, whereas the difference between the Victim and Bystander conditions was not statistically significant. This is reflected by looking at means across scenarios: of the six total scenarios, moral condemnation was greater on five scenarios in the Agent condition, but greater only on three scenarios in the Victim condition, as compared to the Bystander condition (see Table 8). Moreover, the difference between the Agent and Victim conditions was marginally significant, $z = 1.72$, $p = .086$, indicating that participants in the Agent condition condemned personal harm more than did participants in the Victim condition.

Table 8. *Moral judgment by condition and scenario.*

Scenarios	*Agent*			*Victim*			*Bystander*		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Military submarine	183	4.17	2.04	178	4.02	1.95	183	3.62	1.98
Lab scientist	182	5.47	1.93	189	5.57	1.74	201	5.49	1.88
Jungle research	187	4.83	1.99	164	5.13	1.78	190	4.23	2.00
Lifeboat	191	5.18	1.78	173	4.68	1.89	203	5.10	1.99
Climbing group	190	5.82	1.61	185	5.74	1.67	188	5.76	1.57
Enemy doctor	184	4.47	2.07	186	3.92	2.02	190	4.21	2.19

As in Experiment 2, the effects of our perspective manipulation were small. In order to contextualize these results, it is helpful to visualize them in comparison to other effects known in the literature. In Figure 14, we display the regression equation coefficients of our perspective manipulations (employing the dummy coding scheme, 1: Agent/Victim, 0: Bystander) as well as those of the relevant z-scored demographic variables (i.e., political orientation and religiosity). This reveals that the effect of agent perspective-taking on judgments in the purity domain is about half the size of a standard deviation shift in political orientation or religiosity. And, the effect of agent perspective-taking on the judgment of moral dilemmas is roughly equivalent to its effect in the purity domain.

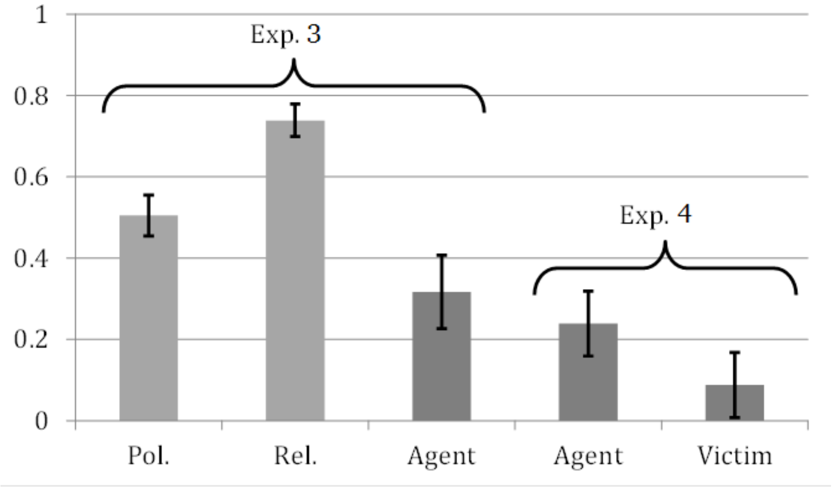


Figure 14. Effect sizes on moral judgment: politics, religiosity and perspective manipulation.

Collapsing across conditions, we replicated the correlation between action focus and moral judgment, *z* = 11.90, *p* < .001. In addition we found a statistically significant but smaller relationship between outcome focus and moral judgment, *z* = 5.43, *p* < .001, indicating that participants who tended to make deontological judgments exhibited greater outcome focus. However, entering both action and outcome foci into the regression model on moral judgment, the effect of outcome focus did not hold, action *z* = 10.64, *p* = .001; outcome *z* = 1.41, *p* = .16. With a larger sample size than Experiment 2 we also observed a modest relationship between deontological moral judgment and ratings of agent perspective-taking, *z* = 5.44, *p* < .001, but not victim perspective-taking, *z* = .71, *p* = .48.

Finally, Experiment 4 replicated the finding that conservatives and religious participants were more agent-focused while liberals were more victim-focused. Participants who adopted the agent's perspective were more conservative (*n* = 382, *M* = 3.62, *SD* = 1.52) than participants who adopted the victim's perspective (*n* = 522, *M* = 3.01, *SD* = 1.48), *t*(928) = 6.17, *p* < .0001. Correlations with ratings of agent and victim perspective-taking and action and outcome foci were replicated (see Table 9).

Table 9. *Political orientation, religiosity and evaluative focus: correlations.*

	Focus		*Perspective-taking*		
	Action	Outcome	Agent	Victim	Relative
Political conservatism	.21 ***	-.08 **	.20 ***	-.13 ***	-.18 ***
Religiosity	.26 ***	.14 ***	.11 ***	.02	-.03

*: *p* < .05; **: *p* < .005; ***: *p* < .0005

4.9. Discussion

Our perspective manipulation yielded effects on the condemnation of moral violations consistent with the agent focus model. When actions were narrated from the perspective of the agent, participants made harsher moral judgments than when they were narrated from a bystander's perspective, suggesting a role for perspective-taking and simulation in moral judgment. This difference held for both harmful actions – like pushing someone to their death in order to save five people – and impure behaviors – like eating one's dead dog. In conjunction with the results of Experiments 1 and 2, these results derive support for the role of unconscious processes of mental simulation in the condemnation of third-party moral violations. If moral judgment were accomplished by categorizing actions according to their causal, intentional and normative properties, but without a role for perspective-taking, we should not have observed a difference between conditions since the action description was matched across conditions. By contrast, the simulation view (as articulated in Chapter 3) can explain the observed difference: the narrative manipulation promoted the adoption of the protagonist's perspective in interpreting the narrated events. Specifically the agent's perspective rendered the evaluator's own aversion to performing harmful actions salient, and this action aversion influenced moral condemnation.

Lastly, Experiment 5 did not draw support for the victim focus model: we found no significant difference between the Victim and Bystander conditions, although there was a trend towards greater condemnation of harm from the victim's perspective. This result suggests that increased empathic concern for the proximal victim drives the condemnation of personal harms weakly at best, and likely contributes less than the alternative mechanism based on action aversion.

4.10. General Discussion

The present study illustrates two related influences on moral judgment: the affective valuation of actions, and the simulation of an agent perspective. In Experiments 1 and 2, individual differences in action focus (but not outcome focus) were correlated with moral judgment, such that greater focus on actions was associated with deontological moral judgment on personal moral dilemmas and greater condemnation of purity violations. Participants who reported making moral judgments by focusing on the action were likely to focus on the aversive character of the actions involved in purity and personal harm violations, and condemn these actions even in the face of consequentialist reasons to condone them. Experiments 3 and 4 demonstrated the causal role of agent perspective-taking in the condemnation of purity and harm violations. When induced to adopt the agent's perspective in our vignettes, participants condemned moral violations significantly more than when induced to adopt a passive bystander's perspective, matching level of detail across conditions.

These findings help to resolve the precise role of affect in the judgment of moral dilemmas that involve a trade-off among lives. The affective response to the footbridge version of the trolley problem, for instance, appears to arise principally not from empathy for the victim being pushed, but rather from an aversion to performing harmful actions, like pushing, that typically bring about harm. As we saw in Chapter 3, moral judgment is sensitive to motor features of the evaluated action, such as whether the agent directly applies muscular force to the victim (Greene et al., 2009). The results reported here suggest that this may be because we unconsciously adopt the agent's perspective and mentally simulate their behavior when judging third parties. This simulation in turn elicits an aversive affective response congruent with *own* performance of the action, which promotes moral condemnation. This might be termed "evaluative simulation", an act of perspective-taking that functions not to describe, predict or explain another's behavior (Gallese & Goldman, 1998; Goldman, 2006; Gordon, 1986), but rather to judge it (Miller & Cushman, 2013; Miller et al., 2013). Prior evidence of the role of own disgust sensitivity in the condemnation of third-party purity violations, along with our present finding that agent perspective-taking moderated this relationship, suggests that processes of evaluative simulation are responsible for condemnation in the domain of purity too. Lastly, further evidence of the interference of visual (Amit & Greene, 2012; Schnall et al., 2008) and olfactory (Inbar et al., 2012) systems with processes of moral judgment bolsters the case for a mechanism of moral evaluation that makes use of sensorimotor representations in condemning the behavior of others.

The mechanism of action-aversion, and its extension to the task of third-party moral judgment, help to explain some familiar yet puzzling aspects of our moral psychology as it is deployed in everyday contexts. Actions that some people find personally aversive yet do not cause obvious direct harm – such as swearing, nudism and homosexual intercourse – may become a target of condemnation. Conversely, actions that are not typically associated with harm and thus fail to elicit a learned action aversion – for instance, using powerful fertilizers for home lawn care, or purchasing sweatshop-manufactured clothing – may be condoned regardless of the harmful outcomes that they ultimately bring about.

An important area for further investigation is the relationship between the evaluative focus on actions versus outcomes and the dyadic theory of moral judgment (Gray, Young & Waytz, 2012). According to the dyadic theory, all moral evaluation involves the perception of a harm-doing agent and a suffering patient, a phenomenon that is termed dyadic completion. There is an evident connection between these agent and patient roles and the emphasis on action versus outcome that we have discussed here. For our present purposes we remain agnostic on whether all moral evaluation *necessarily* involves the perception of both dyadic roles; however, our research does suggest that there may be individual differences in the extent to which an agent's action, versus its outcomes to patients, are the primary target of focus during moral evaluation.

One limitation of the present study is the small-to-medium effect sizes reported throughout. Even in the examination of purity violations, where the conceptual case for the role of agent focus is strong, we did not observe large effects on moral judgment. This weakness may owe to our methodological approach. For example, we sought to measure individual differences in action and outcome foci, processes arguably at a cognitive-attentional level, by employing a psychometric scale perhaps better suited to capture higher-level attitudinal and personality differences. That is, the items on the self-regulation scale express beliefs and attitudes about morality that may signal an underlying action or outcome focus, but they do not directly target the lower-level processes of interest. This methodological "distance" from the target psychological variables may have reduced the validity and accuracy of our indices of action and outcome foci and (if the theoretical framework is correct) could be responsible for the modest effect sizes we observed. Our perspective manipulation similarly did not yield large effects. One reason for this may be the lack of time pressure during exposure to the vignettes, giving participants the opportunity to counteract the manipulation by mentally "shifting out" of the induced perspective.

Additionally, in contrast to the significant effect of manipulating perspective-taking through narrative focus, perspective-taking *via self report* did not consistently predict moral judgment. We have suggested that the relevant processes of agent and victim foci may operate automatically and unconsciously (Bargh & Chartrand, 1999; Hauser et al., 2007), but this hypothesis awaits systematic investigation.

Still, the present study offers some promising first steps in establishing the source of aversive affect in moral judgment. Proponents of the affective revolution in moral psychology have largely taken empathy to be the basis of moral judgment, at least in the harm domain. These studies highlight an alternative account: Moral condemnation largely depends upon an intrinsic aversion to performing harmful actions extended to third parties through an automatic process of evaluative simulation. Outcome-focused affect such as empathy may play a key role in the developmental acquisition of action aversion (Blair, 1995; Blair et al., 1997), but apparently plays a relatively weaker role in generating the widespread condemnation of welfare sacrifices in personal dilemmas. At the level of psychological processing, the analogy to the purity domain seems fruitful: our attitudes concerning the categorical immorality of eating pet dogs and pushing people off bridges depend to a surprising extent on imagining how disturbing it would be to do those things ourselves.

Chapter 5. The influence of evaluative focus on moral and political attitudes

5.1. Introduction

It is patently clear that political liberals and conservatives disagree vehemently on issues related to morality, as evinced both on questions of public policy and in laboratory investigations (Graham et al., in press; Inbar et al., 2009a; Piazza & Sousa, 2013). It is a core ambition of research in political psychology to understand those disagreements in terms of broader cognitive structures.

The previous chapter demonstrated replicable differences in agent and victim foci across the political spectrum. Specifically, conservatives tended to demonstrate an evaluative focus on the intrinsic value of actions, while liberals tended to focus on the action's expected outcomes. So, in this chapter I ask whether the aforementioned difference in evaluative focus might help to explain the patent moral disagreement between conservatives and liberals. However, first I shall briefly review past achievements in our theoretical grasp of the psychological basis of ideological differences.

One successful approach, Moral Foundations Theory, highlights differences in the *content* of liberal and conservative moral codes (Graham, Haidt & Nosek, 2009; Haidt & Graham, 2007; Haidt & Joseph, 2004). Building on earlier attempts to aggregate diverse moral norms into a superordinate structure of basic concerns (Shweder, Much, Mahapatra, & Park, 1997; Haidt & Joseph, 2004), Haidt and Graham (2007) proposed a taxonomy of five foundations that together form the basis of morality across cultures. The foundations of *harm* and *fairness* encompass norms that proscribe harm to others and unjust behavior, and are suggested to relate to the evolution of the mammalian attachment system, empathy and mind-reading, and the development of reciprocal altruism. Together these domains are referred to as the "individualizing" foundations since they function primarily to protect the individual's wellbeing. The *loyalty* and *authority* foundations are comprised of norms that establish societal order and maintain intra-group relations e.g., norms about loyalty to superiors, patriotic duties, respect for elders, as well as govern attitudes towards the outgroup. These foundations are proposed to be closely linked to the evolution of small-scale, hunter-gatherer and tribal communities and their hierarchical structures. Finally, the *purity* foundation contains a wide range of norms concerning food, hygiene, proper use of the body, including sexual behavior, and religious mandates about transcending carnal and animal impulses, and cultivating a spiritual sense. These norms evolved originally to deter the ingestion of contaminant substances, but now serve a broader purpose of establishing standards of decency and propriety. The loyalty, authority, and purity domains are sometimes referred to as the "binding" foundations, suggesting that their proximate social function is to bind groups together rather than to directly protect the individual. It is argued that these five moral foundations are innate and can be found in every moral code across cultures, instantiated by different sets of substantive norms in each social context (Haidt & Joseph, 2004).

In a large-scale study, Graham and colleagues (2009) found that across different cultures and nationalities people who identify with the political left place slightly greater importance on the values of care and fairness, while people who identify with the political right place substantially greater value on the virtues of loyalty toward one's ingroup, respect for authority, and upholding sanctity and bodily purity. An attractive feature of this result is that it can easily explain what is perhaps the most obvious aspect of the left-right political spectrum: disagreement on matters of policy. Questions of the proper scope of military power, immigration, and the role of religion in public life, for instance, clearly can be linked to a theory of divergent moral foundations.

However, while it explains a great deal about political disagreements over morality, this *content*-based schema neglects a class of disagreements that arise in the face of welfare trade-offs. Some experimental evidence indicates that conservatives express greater opposition to directly harming one person in order to save five others in trolley-type dilemmas (Graham et al., in prep.; Piazza & Sousa, 2013), and we replicate this finding in Analysis 1. In addition, a wealth of evidence – discussed in greater detail in previous chapters – indicates that these condemnatory responses to moral dilemmas are motivated by an aversion to harmful action (Bartels & Pizarro, 2011; Greene et al., 2001, 2004, 2009; Koenigs et al., 2007; Miller et al., 2013).

These differences travel beyond the trolley tracks, and are illustrated in salient policy disputes. For instance, the issue of *voluntary euthanasia*, i.e., whether it is right for a physician to assist a terminally-ill patient in fulfilling their desire to die, has fueled a thorny debate in bioethics and philosophy (Foot, 1977; Kass, 1989; Rachels, 1986; Singer, 1995) and consistently drives a wedge between secular liberals and religious conservatives on major opinion polls, with liberals favoring more permissive policies while conservatives favor more restrictive policies (Procon.Org, 2011). Here again, the conservative position seems motivated by a greater, not lesser, moral concern with harmful action.

The trouble, of course, is that – according to the theory of moral foundations – conservatives are supposed to care slightly less about harm than liberals, or certainly no more (Graham et al., 2009). Why, then, should they show increased condemnation of harm in the context of abortion and euthanasia, and also for trolley-type cases? A closer look at the structure of these dilemmas reveals a potential explanation. In each case, a disturbing action is balanced against welfare interests: the welfare of the mother; the suffering of the terminally ill; the five workers on the track. This introduces the possibility that the corresponding liberal and conservative moral views on these issues can be understood in terms of their divergent evaluative foci on the intrinsic moral status of actions viewed in isolation (conservative) versus the expected value of outcomes aggregately (liberal).

Three existing lines of evidence support this structural hypothesis. First, the structural hypothesis dovetails with a recent finding that increased religiosity – which is, of course, strongly correlated with social conservatism (Olson & Green, 2006) – leads to a preference for "rule-based" (i.e. deontological) versus outcome-based moral judgment (Piazza, 2012). Second, a structural approach might help to explain conservatives' and liberals' contrasting approaches to harmless taboos, such as kissing a sibling, peeing in public or eating one's dead pet (Inbar et al., 2009a). As we saw in Chapter 4, the extent which participants focused on actions (but not outcomes) predicted their condemnation of purity domain violations. Moreover, through an induction of the agent's perspective led to increased condemnation (relative to the bystander's perspective), and this difference was attributable presumably to the increased salience of the disgusting action via simulation. Third, the evaluative focus perspective is linked with prior differences in the thinking styles of liberals and conservatives, a point which merits more detailed attention below.

5.1.1. Cognitive style and politics

To identify moral disagreements between liberals and conservatives is only a first step towards the ultimate goal of explaining how underlying cognitive differences shape these moral disagreements. On this question, meager progress has been made. One exception is Janoff-Bulman and colleagues' (2008) linkage of observed differences in approach versus avoidance motivation among liberals and conservatives to their concern with proscriptive ("thou shalt not") versus prescriptive ("thou shalt") norms.

The complementary focus of this chapter is to establish a link between cognitive style on the one hand, and the moral foundations identified by Haidt, Graham, and colleagues on the other. Past studies show that conservatives exhibit a greater need for cognitive closure (Chirumbolo, 2002; Jost, Kruglanski, & Simon, 1999; Webster & Kruglanski, 1994), and are more intolerant of ambiguity (Frenkel-Brunswik, 1948) and avoidant of uncertainty (Wilson, 1973) than are liberals. Moreover, recent studies have demonstrated the longitudinal effects of an intuitive cognitive profile on conservative ideology (Hodson & Busseri, 2012) and on the related construct of right-wing authoritarianism (Heaven, Ciarrochi, & Leeson, 2011). Finally, even the short-term induction of an intuitive thinking style yields a boost in self-reported conservatism and conservative views (Eidelman, Crandall, Goodman, & Blanchar, 2012). In three experiments, Eidelman and colleagues observed that political conservatism increased with blood-alcohol content, and that participants displayed more conservative attitudes under cognitive load, and greater endorsement of conservative terms and rejection of liberal terms under time pressure (see also Shenhav, Rand, & Greene, 2012). Collectively, these studies indicate that a conservative political orientation is associated with – and moreover, causally dependent on – a preference for simpler, definitive, and intuitive styles of thinking. How might these thinking styles explain different moral attitudes? In other words, why might the conservative preference for intuition and epistemic certainty lead to stricter sexual norms or stronger favoritism towards the ingroup, for instance, while a preference for reflection leads to a selective concern with harm and fairness?

Through the lens of a content-based schema, it is not obvious why this should be. Adopting a structural approach, however, these connections are more readily explained. In the previous chapter, I presented evidence that people vary in the degree to which they engage two different approaches to moral judgment. One approach, which we referred to as victim focus, involves an assessment of whether the *expected outcomes* of a behavior are positive or negative. Attention is directed at the moral patients and the outcomes likely to befall them. An alternative approach to making moral judgments, which we labeled agent focus, involves a focus not on the outcomes of the action, but rather on the intrinsic moral status of the action.

We considered the example of a teacher slapping a child in the previous chapter. Through an evaluative focus on actions, one's categorical aversion to slapping children may give rise to the simple judgment that the teacher's action is inherently wrong. This approach likely depends on the acquisition of a valenced attitude toward the action itself, and a spontaneous simulation of the affective experience of performing the action, which gives rise to a moral evaluation.[17]

By contrast, an evaluative focus on outcomes implicates more demanding cognitive processes, insofar as the teacher's action may have a multitude of potential outcomes. Beyond the child's immediate pain and humiliation, it may also discourage the child's undesirable behavior, decrease his self-esteem, deteriorate his relationship with the teacher, and may also indirectly affect others—for instance, it may influence the child to behave violently towards his peers. Moreover, there is uncertainty associated with each of these outcomes: they may or may not happen.

So, evaluating actions based upon the intrinsic value representations that they immediately invoke—does this feel right, or wrong?—provides a heuristic approach to moral judgment that can potentially yield quick and definitive answers (Slovic, Finucane, Peters, & MacGregor, 2002). Whereas a person (in the position of a moral agent) may readily feel that it is wrong to cheat, steal or lie and experience relative certainty in this judgment, in many circumstances identifying and assessing the precise

[17] There are numerous ways these action-based affective responses might arise: some may be largely innate (see Greene et al., 2009; Lieberman & Lobel, 2012), others acquired through cultural learning of taboos and symbolic behavior (Baron & Spranca, 1997), and so on. In addition, Blair's (1995; see also Blair et al., 1997) model of the *violence inhibition mechanism* offers a specific account of the acquisition of many action-based aversions in the harm domain. This issue is treated more thoroughly in Chapter 6.

outcomes of these actions (to the victims and beneficiaries) involves a more exhaustive analysis of multiple subsequent events, each of which is uncertain. Therefore, the simplicity afforded by the action-based approach may be favored, and moreover in most circumstances it will lead to the same judgment as the outcome-based approach, like many heuristics (Kahneman & Frederick, 2002).[18]

This observation suggests a potential role for cognitive style in moral psychology. Individuals who enjoy thinking and engage deep reflection in order to form beliefs should be more likely to make moral evaluations through the cognitively taxing exercise of patient focus. By contrast, people who dislike thinking deeply and prefer to form beliefs in the wake of minimally sufficient evidence may accordingly prefer the strategy of agent focus. This suggests, since liberals and conservatives differ on this psychological dimension, that conservatives should tend to exhibit greater agent focus than liberals, while liberals should tend to exhibit greater patient focus than conservatives (as observed in Chapter 4).

Motivated by these several lines of evidence, the present study aims to illustrate the value of characterizing the basis of moral disagreements between liberals and conservatives not only in terms of moral foundations, but also in terms of the structural division between action-based and outcome-based value representation. Specifically, we predict that agent focus gives rise to a typically conservative moral profile, while patient focus gives rise to a typically liberal moral profile. In addition, this exercise will help to bridge the gap between two areas of past research, asking how fundamental differences in *cognitive style* between liberals and conservatives yield the corresponding differences in moral values described by *moral foundations theory*. In other words, what is the

[18] Indeed, experimental evidence demonstrates that the categorical condemnation of actions, including lying (Greene & Paxton, 2009) and cheating (Rand, Greene & Nowak, 2012), is relatively automatic and cognitively undemanding, while an outcome-sensitive evaluation requires greater cognitive resources (Bartels, 2008; Greene et al., 2008; Paxton, Ungar, & Greene, 2011).

nature of the psychological processes of moral evaluation that give rise to typically conservative and liberal moral profiles?

5.1.2. Outline

In Analysis 1, we replicate a finding that conservatives tend to condemn personal harm for the greater good more than do liberals. This finding cannot be straightforwardly captured by content-based schemas, like the extant Moral Foundations Theory. We propose that this result may owe to differences in the structural approaches of liberals and conservatives. So in Analysis 2 we examine the relationship between evaluative foci and demographics, paying special attention to religiosity and political orientation. Next, in Analysis 3, we examine whether agent and patient foci are related to the established differences in cognitive style between liberals and conservatives. Finally, in Analyses 4 and 5, we examine whether the differences in evaluative foci are at the root of distinct systems of moral foundations, and derivatively, may account for the disagreement about the relevance of binding foundations.

5.1.3. General methods

In this study, we report data that were collected using two online platforms. Participants from both platforms were directed to an online survey where, after a brief introduction detailing the purpose and nature of the study, they completed the Act/Impact Morality Scale (see Experiment 1 in Chapter 4). In each experiment, participants completed one or more additional measures. At the end of the testing session, all participants provided optional demographic information.

Demographics

Participants indicated their *age* on a 10-year interval U.S Census scale (Under 15 years: 2.6%; 15 to 24 years: 54.8%; 25 to 34 years: 21.2%; 35 to 44 years: 11.7%; 45 to 54 years: 8.0%; 55 to 64 years: 3.2%; 65 years and over: 1.1%), *gender* (53.0% females, 47.0% males), *educational attainment* (Less than high school: 11.6%; High school/GED: 19.7%; Some college: 23.9%; 2-year college degree: 8.0%; 4-year college degree: 21.0%; Master’s degree: 12.2%; Doctoral Degree: 2.0%; Professional degree, e.g. JD, MD: 1.7%), *religiosity* (“How religious are you?”; 1 Not at all: 33.7%; 2: 13.1%; 3: 6.7%; 4 Somewhat: 17.0%; 5: 11.5%; 6: 9.7%; 7 Very: 8.4%), and several measures related to political orientation. Participants reported their political orientation on *social issues* (“When it comes to social issues, how liberal or conservative are you?”), *fiscal issues* (“When it comes to economic issues, how liberal or conservative are you?”) and the political orientation that prevailed throughout their development, which we term *familial political orientation* (“How would you describe the political environment that you grew up in, including your family, friends and community?”; see Table 10). Finally, participants indicated the amount of formal education they had received in the subjects of *philosophy* (Very little or none: 53.5%; Some (e.g. one or two courses): 37.1%; A lot (e.g. three or more courses): 9.4%) and *psychology* (Very little or none: 44.4%; Some (e.g. one or two courses): 40.6%; A lot (e.g. three or more courses): 15.1%).

Table 10. *Political orientation: distribution.*

	VL			*M*			*VC*
	1	2	3	4	5	6	7
Social issues	257	305	199	323	137	53	49
(*n* = 1323)	19.4%	23.1%	15.0%	24.4%	10.4%	4.0%	3.6%
Fiscal issues	187	228	203	401	168	77	71
(*n* = 1335)	14.0%	17.1%	15.2%	30.0%	12.6%	5.8%	5.3%
Familial orientation	31	55	63	148	69	56	44
(*n* = 466)	6.65%	11.8%	13.5%	31.8%	14.8%	12.0%	9.4%

Notes. VL: Very liberal; M: Moderate; VC: Very conservative.

5.2. Analysis 1: Focus- versus content-based accounts of moral disagreement.

Recent reports indicate that political conservatives tend to judge five-versus-one welfare trade-offs more harshly in trolley-type contexts (Graham et al., in prep.; Piazza & Sousa, 2013), and we replicate this finding in Analysis 1. We also examined participants' concern for each of the five moral foundations, of harm, fairness, ingroup loyalty, authority and purity.

The most straightforward application of a content-based approach to moral judgment would predict that the condemnation of harmful action in trolley-type dilemmas will correlate strongly with ratings of the harm foundation, but weakly with the remaining foundations, if at all.

A structural approach to moral judgment, in contrast, predicts a strong correlation between the judgment of trolley-type dilemmas and the purity foundation. The moral condemnation of purity-related issues depends not on a consideration of the action's outcomes (Haidt et al., 2000), but rather on an aversion to the action (Lieberman & Lobel, 2013; see also Chapter 4). Similarly, ample evidence indicates that the moral condemnation of personal harm in trolley-type dilemmas rests upon an affective prohibition of actions that directly harm another (Greene et al., 2001 2004; Koenigs et al., 2007; Miller et al., 2013).

5.2.1. Methods

144 participants (81 females) voluntarily logged on to the Moral Sense Test (MST). All research was conducted under the guidelines of the institutional review boards at Harvard University and Brown University. This study was specifically approved by the named institutional review boards, and written consent was obtained from all research participants.

Participants read a brief introduction to the study and then viewed three high-conflict personal moral dilemmas in a random order (Trolley, Crying Baby, and

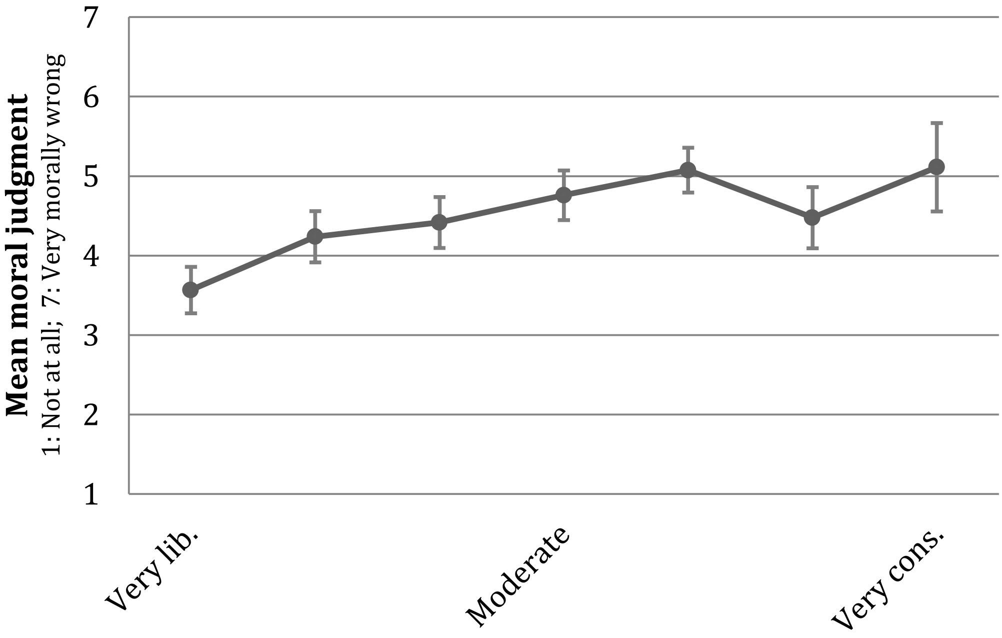
Lifeboat adapted from Greene et al., 2001; see *Appendix E*). These hypothetical dilemmas describe a person who faces the choice whether to kill one person in order to save a greater number, or instead to do nothing and allow the greater number of people to die. In each context, the person chose to sacrifice the one in order to save the greater number, and participants made moral judgments of their behavior on a 7-point Likert scale from 1: "Not morally wrong at all" to 7: "Very morally wrong". These particular scenarios were classified as "high-conflict" in Koenigs et al., 2007, based on long reaction times and high disagreement about the moral permissibility of the behavior. As such, the mean response across our three dilemmas should serve as a reliable index of moral attitudes towards welfare trade-offs involving human death.

Participants also completed the Moral Foundations Questionnaire (MFQ; Graham et al., 2009), a 30-item scale that indexes participants' level of concern with each of the five moral foundations. In the first part, participants rate how relevant various concerns are to them when thinking about moral issues, e.g., "Whether or not someone acted unfairly" (Fairness) and "Whether or not an action caused chaos or disorder" (Authority). In the second part, participants express their agreement or disagreement with a set of items that reflect the relevance of a particular moral foundation, e.g., "One of the worst things a person could do is hurt a defenseless animal" (Harm), "It is more important to be a team player than to express oneself" (Loyalty), and "Chastity is an important and valuable virtue" (Purity). A score for each foundation is derived by averaging across the three corresponding items in the first part and the three corresponding items in the second part. Harm and fairness scores can be averaged into a single metric of concern for the *individualizing* foundations; and similarly loyalty, authority and purity scores are sometimes averaged into a single metric of relevance of the *binding* foundations. The MFQ contains two catch items, and participants who demonstrated inattentiveness on these two items were excluded from the analyses (above 2 on “Whether or not someone was good at math”, and below 5 on “It is better to do good than to do bad”, both on a 6-point scale).

The order of presentation of the moral dilemmas block and the MFQ was counterbalanced across participants. At the end of the study, participants optionally provided some demographic information. The MFQ contains two catch items, and 21 participants who demonstrated inattentiveness on these items were excluded from the following analyses.

5.2.2. Results

We found that moral judgments of trolley-type dilemmas correlated with political orientation on social issues, *r* = .280, *p* < .002, and fiscal issues, *r* = .233, *p* < .01, and marginally with familial political orientation, *r* = .171, *p* = .06, indicating that conservatives tended to condemn personal harm in welfare trade-offs more than did liberals (see Figure 15).[19]



[19] We replicated the finding, presented in Chapter 4,that condemnation of personal harm correlated with action focus, *r* = .261, *p* = .004, but not outcome focus, *r* = -.02, *p* > .8, indicating that the more participants tended to focus on the character of actions, the more they condemned third-party utilitarian trade-offs in personal moral dilemmas.

Figure 15. Moral judgment by political orientation. (Very lib.: Very liberal; Very cons.: Very conservative.)

Next, we turned to the relationship between condemnation of personal harm and relevance of each of the moral foundations. As predicted by the content-based approach, condemnation of personal harm correlated with relevance of the harm foundation, r = .365, p < .0001. However, as predicted by evaluative focus alone, we found correlations between condemnation of personal harm and all three binding foundations; loyalty r = .290, p < .002, authority r = .336, p < .0001, purity r = .434, p < .0001, indicating that the more participants condemned the welfare trade-off in trolley-type dilemmas, the more they cared about loyalty, authority and particularly about purity.

5.2.3. Discussion

Analysis 1 demonstrated that that social conservatives judged welfare trade-offs implicating personal harm more harshly than did liberals. This difference, not captured by moral foundations theory, is easily explained by appealing to evaluative foci. From an evaluative focus perspective, the presence of both action and outcome sources of affect in harm violations introduces the possibility of moral conflict and disagreement. By focusing on the moral agent, conservatives tend to condemn the welfare trade-off on the basis of an affective response to the agent's harmful action (e.g., pushing, shooting a gun, smothering one's child; see also Miller et al., 2013). Liberals, on the other hand, focus on the moral patients and judge the welfare trade-off relatively more permissible on the basis of a concern for the welfare of the patients in the moral interaction.

This philosophical thought experiment provides some purchase on contemporary political divisions. Consider, for instance, euthanasia or abortion. When we simulate these behaviors from the agent's perspective, and focus on the performance of a harmful action, action aversion promotes the categorical condemnation of the behavior (Miller et al., 2013). By contrast, when we jointly consider the expected outcomes of the action, i.e., the victim’s immediate pain but also a broader range of probabilistic, “distal” outcomes associated to the welfare of other patients, these concerns may outweigh the aversion to performing harm to a proximal victim and guide approval of associated policies.

In the second part of this study, we found that increased reliance on the harm foundation predicted deontological moral judgment of dilemmas, consistent with either content-based or structure-based schemas. However, increased sensitivity to purity concerns also predicted deontological moral judgment of dilemmas and, if anything, this latter correlation appeared to be the stronger of the two. This result seemingly favors the structural hypothesis. Still, we must ask whether these results can be understood in terms of a content-based approach alone. It is possible, for instance, that these moral dilemmas are construed as purity-related issues: e.g., “Human life is sacred, and not to be defiled by its profane sacrifice for greater welfare”. Such language is often invoked in opposition to abortion and euthanasia, although we may detect a whiff of *post hoc* rationalization about it.

Therefore, though moral foundations theory captures the disagreements between liberals and conservatives that surround transgressions of the binding foundations, the theory neglects another category of issues, concerning the permissibility of welfare trade-offs, that generate divisive moral disagreements. Analysis 2 employs a more direct measure of evaluative focus in moral judgment across the political spectrum.

5.3. Analysis 2: Demographic predictors of evaluative focus.

In Experiment 2 we ask participants whether they adopt a structural approach to moral judgment that places greater emphasis on the feelings associated with actions intrinsically, or instead on the outcomes that those actions are likely to cause. Our prediction is that conservatives will emphasize the former set of concerns, while liberals will emphasize the latter.

5.3.1. Method

493 participants (276 females) logged on to the MST, where they first read a brief introduction to the study and provided their written consent. Participants then completed our measures of evaluative focus (see Chapter 4), and then voluntarily provided demographic information.

5.3.2. Results

As predicted, social conservatives demonstrated greater action focus, $r = .212$, $p < .0001$, and lesser outcome focus, $r = -.141$, $p < .003$, than did social liberals. In a multiple regression model, these were unique effects on social political orientation, AF $\beta = .267$, $p < .001$, OF $\beta = -.211$, $p < .001$. With our measures of third-party evaluation, we obtained the corresponding pattern of results. Evaluative simulation (via focus on the moral agent) was associated with a conservative stance on social issues, $r = .158$, $p < .001$, while outcome assessment (via focus on the moral patients) was associated with liberal views on social issues, $r = -.193$, $p < .0001$, and these were independent effects in a multiple regression model, ES $\beta = .127$, $p < .01$, OA $\beta = -.170$, $p < .001$. Finally, we observed that participants who endorsed evaluative simulation were more socially, $t(463) = 4.02$, $p < .0001$, and fiscally, $t(465) = 2.40$, $p < .02$, conservative than participants who endorsed outcome assessment. With views on fiscal policy, we observed correlations with our measures of action focus, $r = .132$, $p < .005$, and evaluative simulation, $r = .120$, $p < .01$, only. Similarly, religiosity correlated only with action focus, $r = .175$, $p < .003$.

In both self-regulation and perspective-taking during third-party evaluation, we found differences along the political spectrum, which were larger for *social*, than for *fiscal*, political orientation. (Figure 16 displays mean values on measures of evaluative focus across the political spectrum.) Other demographic measures, including academic instruction in philosophy and psychology, gender, religiosity, and education, correlated also with evaluative foci in diverse ways, though these effects were smaller and less stable across our related measures than the effects on political orientation. See Supplementary Analysis 1 for correlations with other demographic variables.

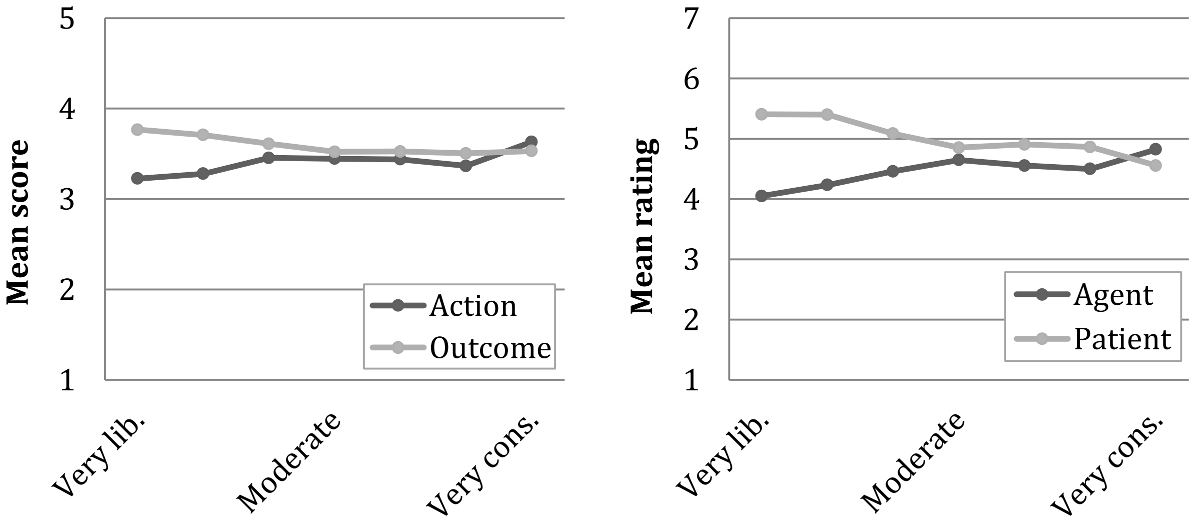


Figure 16. Action and outcome focus (left) and ratings of agent and patient perspective-taking (right) by social political orientation. (Very lib.: Very liberal; Very cons.: Very conservative.)

5.3.3. Discussion

Social conservatives demonstrated an evaluative focus on actions and tended to endorse evaluative simulation, while social liberals demonstrated greater evaluative focus on outcomes and tended to favor outcome assessment. Political orientation on fiscal issues did not yield as clear systematic effects on evaluative focus and, surprisingly, neither did religiosity (but see Piazza, 2012). With familial conservatism, we found no relationships in the first place; this result provides suggestive evidence that evaluative focus is not the result of a liberal or conservative upbringing, but rather derives from non-heritable factors that may subsequently shape political orientation.

These demographic variables correlated with our measure of action focus, but the overall size of the effects was stronger with social political orientation.

5.4. Analysis 3: The influence of cognitive style on moral evaluation.

In Analysis 3, we employ two differentiated dimensions of cognitive style, *need for cognition* (Cacioppo, Petty & Kao, 1984) and *need for closure* (Webster & Kruglanksi, 1994), in order to determine the relationship between participants' structural approach to moral judgment and their cognitive style (intuitive versus deliberative). This is motivated by a straightforward observation: To make moral judgments by assessing how an action "feels" plausibly depends on System 1 processing (akin to the *affect heuristic* in Slovic et al., 2002), while to make moral judgments by assessing the likely outcomes of an action requires a System 2-based search over probabilistic causal models. These two approaches to decision-making have been formalized in computational terms (Sutton & Barto 1999; Dayan & Niv, 2008; Daw & Shohamy, 2008) and signatures of both approaches are evident behaviorally (Otto, Gershman, Markman & Daw, 2013; Gershman, Markman & Otto, 2012) and neurally (Gläscher, Daw, Dayan, & O'Doherty, 2010). Altogether, these studies indicate that "action-based" evaluation operates in a more automatic fashion, whereas "outcome-based" evaluation depends more upon mechanisms of cognitive control (Crockett, 2013; Cushman, 2013). We therefore predict that participants with an intuitive cognitive style will exhibit an evaluative focus on actions, while participants with a reflective cognitive style will exhibit an evaluative focus on outcomes.

5.4.1. Methods

236 participants (128 females) were recruited via Amazon's Mechanical Turk (www.mturk.com), an online labor market where short incentivized tasks are completed by workers worldwide. Several studies demonstrate that Mechanical Turk provides a

diverse sample for affordable and high-quality data collection in Web-based behavioral paradigms (Buhrmester, Kwang, & Gosling, 2011; Horton, Rand, & Zeckhauser, 2011).

Participants provided their written consent and then completed two widely-used measures of cognitive style – the 18-item *Need for Cognition Short Form* (Cacioppo et al., 1984), and the 42-item *Need for Closure Scale* (Webster & Kruglanski, 1994) – along with our measures of evaluative focus, in a randomized order to avoid the influence of systematic order effects upon our data.

The 18-item *Need for Cognition – Short Form* is a widely-used and reliable measure of need for cognition, or "an individual's tendency to engage in and enjoy effortful cognitive endeavors" (Cacioppo et al., 1984, p. 306). People who display a high need for cognition find thinking enjoyable and engage in reflection frequently, whereas people who display low need for cognition find thinking tedious and unappealing and engage thinking only when it is properly incentivized. Example items are "I really enjoy a task that involves coming up with new solutions to problems" and "I only think as hard as I have to" (reverse-scored). The sum score of all 18-items, ranging from 18 to 90, was computed as a participant's need for cognition index.

The 42-item Need for Closure Scale, designed to assess individual differences in the desire to possess knowledge (on a given topic), rather than enduring ambiguity or indecision (Kruglanski, Webster, & Klem, 1993; Webster & Kruglanski, 1994). People who are high in need for closure tend to seek definitive knowledge and form conclusive beliefs with minimally sufficient evidence. This trait is known to give rise to a preference for order and routine and an intolerance of ambiguity and unpredictability. By contrast, people who are low in need for closure are more likely to endure ambiguity and uncertainty, and hold higher standards for the formation of beliefs. The need for closure construct is composed of five subscales: order (e.g. "I hate to change my plans at the last minute"), predictability (e.g. "I enjoy the uncertainty of going into a new situation without knowing what might happen,” reverse-scored), decisiveness (e.g. “When faced with a problem I usually see the one best solution very quickly”), ambiguity (e.g. “I feel uncomfortable when someone's meaning or intention is unclear to me”), and closed-mindedness (e.g. “When considering most conflict situations, I can usually see how both sides could be right,” reverse-scored). The sum score of all the items, ranging from 42 to 252, was computed as a participant’s need for closure index. At the end of the testing session, participants optionally provided basic demographic information.

5.4.2. Results

We replicated previous findings demonstrating a relationship between cognitive style and political orientation (Jost et al., 1999): specifically, high need for closure and low need for cognition were associated with social conservatism, closure $r = .193$, $p < .008$, cognition $r = -.223$, $p = .002$, but bore no relationship to fiscal conservatism, *p*s > .5. Need for cognition and need for closure were also negatively correlated with each other, $r = -.290$, $p < .0001$.

As predicted, action focus correlated positively with need for closure, $r = .330$, $p < .0001$, and negatively with need for cognition, $r = -.224$, $p < .002$, confirming that action focus was associated with an intuitive cognitive style. By contrast, outcome focus did not correlate with either need for cognition or need for closure, *p*s > .5 (see Figure 17). Controlling for political orientation in separate multiple regression models, the effects of action focus on the indices of cognitive style remained significant, cognition $\beta = -.223$, $p < .002$, closure $\beta = .316$, $p < .001$.

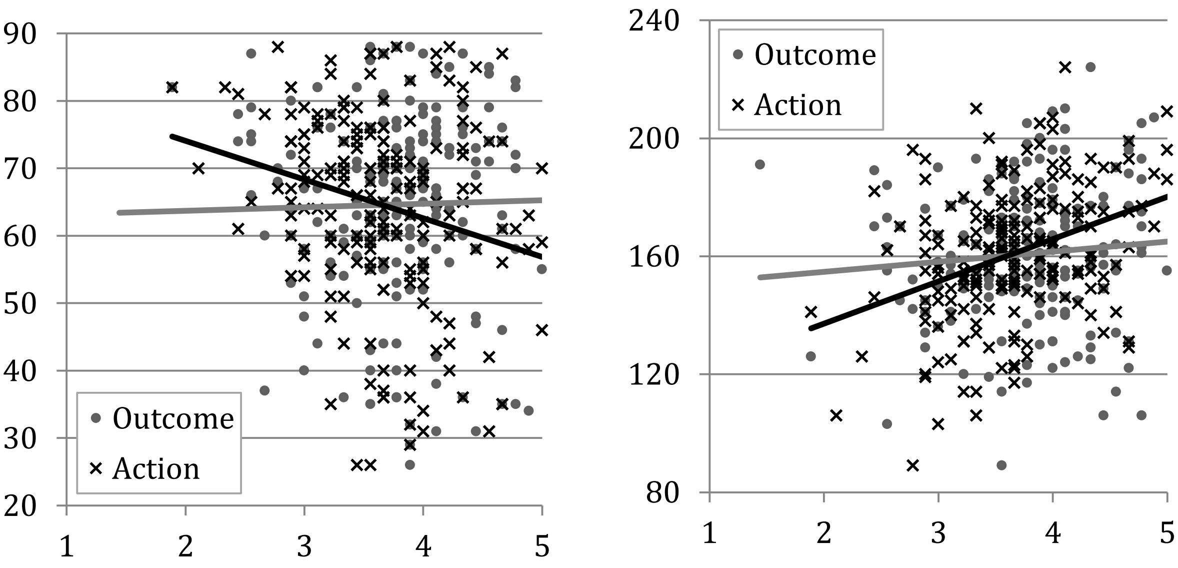


Figure 17. Need for cognition (left) and need for cognitive closure (right) foundations by action and outcome focus.

There were no effects of need for cognition or need for closure on the endorsement measure or relative rating of agent vs. patient perspective-taking, all *p*s > .15, suggesting that perspective-taking in third-party evaluation is not associated to cognitive style. Examining the relationships of cognitive style to the independent ratings of agent and patient perspective-taking, once again we found little evidence for systematic relationships: Agent perspective-taking correlated with neither need for closure nor need for cognition, *p*s > .3, while patient perspective-taking correlated with need for cognition, *r* = .187, *p* = .008, but not need for closure, *p* > .3.

5.4.3. Discussion

We found a pattern of results consistent with the view that an intuitive cognitive style supports an agent focus while a reflective cognitive style supports a patient focus in moral judgment. However, the precise set of results was rather more nuanced: We found that intuitive cognitive style (indexed by a low need for cognition and a high need for closure) predicted social political orientation and, in a multiple regression model, rendered the effect of action focus on political orientation non-significant. This result has some intuitive plausibility. Cognitive style shapes political orientation in a variety of ways, many of which do not have a moral dimension. For instance, the preference for conformity and tradition, appreciating simple and intuitively appealing arguments, or being suspicious of science and higher education are elements of intuitive thinking that may promote conservatism in ways that evaluative focus could not explain. The results of the multiple regression analysis are consistent with this interpretation: cognitive style influences politics via numerous routes, only some of which implicate evaluative focus, and thus remains the stronger predictor of political orientation.

By contrast, with third-party evaluation, we found minimal relationships between cognitive style and either agent or patient perspective-taking, suggesting that differences in perspective-taking along the political spectrum are not principally due to differences in cognitive style. One explanation for this might be that whether we take our aversions to particular actions into account when deciding what is right for *us* to do is largely due to a general reliance on intuition. But, given those feelings, whether we employ them to judge others through evaluative simulation is not influenced by cognitive style, but rather a *sui generis* approach to third-party moral judgment.

The results of Analysis 3 demonstrate a relation between cognitive style and *social* political orientation, but no relation with *fiscal* political orientation. Together with the results of Analysis 2 (where differences in evaluative foci where stronger with social than fiscal political orientation too), this suggests that the phenomenon of interest, i.e., how cognitive style gives rise to distinct approaches to morality among liberals vs. conservatives, primarily implicates the social dimension of political ideology. Thus, in further analyses, we focus on the relations that obtain with social political orientation.

Finally, we found that cognitive style was strongly related to our measures of action focus, but weakly if at all related to our measures of outcome focus. (In contrast, outcome-focus items did correlate with political orientation in Analysis 2). We did not anticipate this result, which deserves further investigation. Possibly, our outcome items index the endorsement of outcome-based concerns in the abstract, which is not cognitively demanding or ambiguous (see Ditto & Liu, 2012). Rather, cognitive demands and ambiguity may arise principally from the process of actually imposing outcome-based moral judgments in practice, especially in cases such as trolley-type dilemmas and victimless crimes where they conflict with action-based moral judgments.

At a broad level, the results of Analysis 3 suggest a potential link between two well-studied aspects of political psychology that have previously been treated independently. On the one hand, political conservatives show a preference for intuitive cognitive styles (Jost et al., 1999). On the other hand, conservatives exhibit heightened concern for the binding moral foundations (Graham et al., 2009; Inbar et al., 2009a). We have already demonstrated a connection between an intuitive cognitive style and the preference for action-based moral evaluation. In Analysis 4, we complete the linkage by demonstrating that a preference for action-based moral evaluation is linked to the adoption of conservative moral values.

5.5. Analysis 4: Evaluative foci predict patterns of moral foundations.

Analysis 4 seeks to examine the relationship between evaluative foci and systems of moral foundations, using the Moral Foundations Questionnaire (MFQ; Graham et al., 2009). It is already known that political liberals primarily rely upon two psychological foundations, the foundations of harm and fairness, whereas political conservatives rely more evenly upon all five psychological foundations. The structural hypothesis provides a possible explanation for these observed differences: Actions that violate the individualizing foundations, such as sticking a pin into the palm of a child or cheating in a game of cards, involve an agent harming one or more patients. By contrast, actions that violate the binding foundations, such as burning one’s national flag or cooking and eating one’s dead pet dog, feature agents performing reprehensible acts that in many instances do not entail harmful outcomes to patients. They of course may do so: insulting a family member, or cursing one’s nation on the radio, for instance, likely bring about some victim distress in each case. But critically, an action need not entail a harmful outcome in order to constitute a violation of the loyalty, authority or purity foundations.

Consequently, we predict a strong relationship between outcome focus and concern with the harm and fairness foundations, sometimes referred to as *individualizing* foundations. We also predicted a strong relationship between action focus and the harm and fairness foundations, insofar as one might be motivated by an aversion to harmful or unfair actions just as much as by the outcomes they produce (Cushman et al., 2012; Miller et al., 2013). For the purity, authority and loyalty foundations—sometimes referred to as *binding* foundations—we predicted a strong relationship with action items alone. Thus, the individuals who exhibit agent focus should demonstrate greater concern for the binding foundations than patient-focused individuals demonstrate.

5.5.1. Methods

571 participants (323 females) voluntarily logged on to the Moral Sense Test, read a brief introduction to the study and provided their written consent. Next, participants completed the assessment of evaluative focus developed in Chapter 4 along with the MFQ. 93 participants who demonstrated inattentiveness through the catch items in the MFQ were excluded from subsequent analyses. At the end of experiment, participants optionally provided demographic information.

5.5.2. Results

As predicted, the extent to which participants exhibited an evaluative focus on actions correlated with their concern for all five foundations, whereas the extent to which they emphasized outcomes correlated only with their concern for individualizing foundations (see Figure 18, and statistical tests presented in Table 11). Next we examined the measures of third-party evaluation and found the corresponding pattern of relations: ratings of evaluative simulation correlated with moralization across all five foundations, whereas ratings of outcome assessment correlated only with moralization of the individualizing foundations (see Table 11).

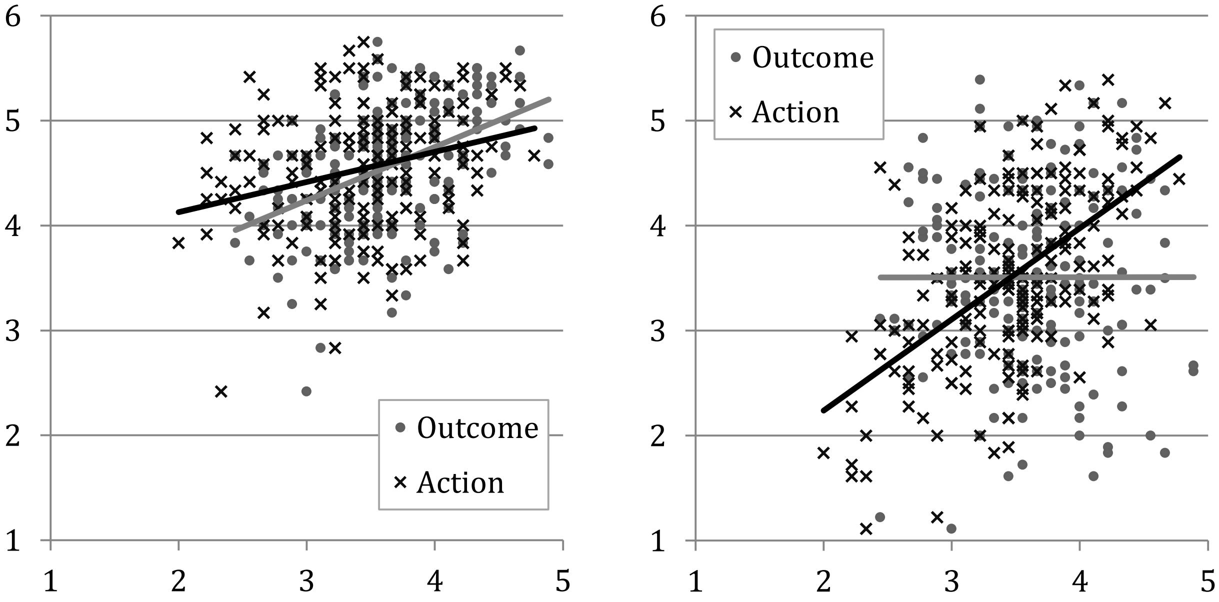


Figure 18. Relevance of the individualizing (left) and binding (right) foundations by action and outcome focus.

Table 11. *Evaluative foci and moral foundations: correlations.*

	Harm	*Fairness*	*Loyalty*	*Authority*	*Purity*
Action focus	.288***	.224***	.278***	.376***	.451***
Outcome focus	.431***	.420***	.075	-.014	.018
Agent perspective-taking	.132**	.102*	.139**	.198***	.230***
Patient perspective-taking	.312***	.322***	-.025	-.053	-.071

Notes. * *p* < .05; ** *p* < .01; *** *p* < .001.

Looking at the dichotomous endorsement measure, we found that participants who reported adopting the agent's perspective in third-party evaluation judged the individualizing foundations as less relevant, and the binding foundations as more relevant, to them than did participants who reported adopting the patient's perspective, individualizing $t(477) = -3.76$, $p < .0005$, binding $t(477) = 2.23$, $p = .026$. This relationship was confirmed on the bipolar relative rating of agent vs. patient perspective-taking: as participants rated agent perspective-taking more favorably than patient perspective-taking, they tended to judge the individualizing foundations as less relevant, $r = .153$, $p < .001$, and the binding foundations as more relevant to their moral judgment, $r = -.150$, $p < .001$.

We then entered our measures of evaluative focus along with social political orientation into separate multiple regressions predicting each of the moral foundations (see Supplementary Analysis 2). Still, we observed significant effects of action focus on all five foundations, $.185 < \beta\text{s} < .371$, *p*s < .001, and of outcome focus on the individualizing foundations, $.333 < \beta\text{s} < .369$, *p*s < .001, after controlling for the effects of political orientation. Similarly, we entered our measures of third-party evaluation along with social political orientation into separate multiple regressions predicting each of the moral foundations (see also Supplementary Analysis 3). Here too we observed independent effects of evaluative simulation on the individualizing foundations, $.162 < \beta\text{s} < .181$, *p*s < .001, and outcome assessment, $.291 < \beta\text{s} < .336$, *p*s < .001, after controlling for political orientation. The effects of evaluative simulation on each of the binding foundations after controlling for political orientation were significant $.098 < \beta\text{s} < .163$, *p*s < .05, though their overall sizes were small. In sum, the effects of participants' evaluative foci on their pattern of moral foundations held even after controlling for political orientation.

In a follow-up study, we confirmed the above findings employing a complementary measure, the Moral Foundations Sacredness Scale (MFSS; Graham et al., 2009). On the MFSS, participants are requested to imagine performing a series of actions. The actions listed on MFSS are moral violations corresponding to each of the five foundations, .e.g. “Kick a dog in the head hard” (Harm), “Renounce your citizenship and become a citizen of another country” (Loyalty), or “Get a blood transfusion of 1 pint of disease-free, compatible blood from a convicted child molester” (Purity). Participants select the appropriate point on an 8-point logarithmic scale, indicating “how much money someone would have to pay [them] (anonymously and secretly) to be willing to do each thing” (1: $0 I’d do it for free, 2: $10, 3: $100, 4: $1000, 5: $10,000, 6: $100,000, 7: $1,000,000, 8: Never for any amount of money). As such, the MFSS provides a visceral counterpart to the more abstract MFQ. A measure of *unwillingness to violate* (or sacredness of) a given foundation is derived by averaging across the four items corresponding to each foundation. Since the MFSS measures the sacredness of each foundation in the participant’s *own* behavior, in these analyses we were primarily interested in the relationships with measures of moral self-regulation. We found that both action and outcome foci correlated with moralization across all five foundations, all *p*s < .01. So, we entered action and outcome foci into separate multiple regression models for each foundation: action focus and outcome focus each explained unique variance in the sacredness of the harm and fairness foundations (all *p*s < .01), whereas only action focus explained unique variance in the loyalty, authority and purity foundations (action focus *p*s < .001; outcome focus *p*s > .1).

5.5.3. Discussion

As predicted, participants’ evaluative focus on actions versus outcomes predicted the moral views of liberals and conservatives across different stimuli pertaining to the moral foundations. We found that an emphasis on actions was associated with a five foundation morality (the conservative moral profile) whereas an emphasis on outcomes was associated with a harm-based, two foundation morality (the liberal moral profile)**.**

5.6. Analysis 5: Building a model of politics, moral foundations and evaluative foci.

We have thus far demonstrated that evaluative foci are related to political orientation and to distinct patterns of moralization of individualizing and binding concerns. In this section, the aim is to determine the causal ordering among these variables. We first review past evidence suggesting that the moral attitudes in fact shape political orientation, and then employ statistical modeling techniques in order to determine the position of evaluative focus within this causal relationship.

A number of findings indicate that the induction of binding concerns causes participants to identify with the political right. In one recent study, it was found that brief exposure to the national flag promotes identification with the conservative party (Carter, Ferguson, & Hassin, 2011). In a similar vein, attendance to Independence Day parades in the United States throughout childhood predicts conservatism during adulthood (Madestam & Yanagizawa-Scott, 2012). These studies can be understood as demonstrating that political conservatism is promoted by the induction of loyalty and authority concerns. Similarly, reminders of purity have been shown to sway participants towards the conservative end of the spectrum (Helzer & Pizarro, 2011). As a prime of purity concerns, participants completed the experiment next to a hand sanitizer as opposed to across the hall from the hand sanitizer in the control group. In two studies, participants who were exposed to the manipulation reported more politically conservative attitudes than did participants in the control group. Might the causal link also run in the opposite direction? That is, does affiliation with the conservative party shape concern for the binding foundations? Although we do not rule out this relation, at least some evidence from our own data suggests otherwise. Specifically participants' familial political orientation did not predict their own attitudes toward the moral foundations, despite the fact that it did predict their own present political orientation.[20] In this section, we perform Sobel-Goodman mediation analyses on the data in Analysis 4 to arbitrate between three models that are consistent with a causal link from moralization of the binding foundations to political conservatism. We then employ path analysis on these data in order to develop a prototypical causal model of evaluative foci, moral foundations and political orientation.

5.6.1. Results

We compare three mediation models with action focus, concern for the binding foundations and political orientation that are compatible with a causal link from concern for the binding foundations to political conservatism. We posit concern for the binding foundations as causally prior to political orientation, and place agent focus – indexed by action focus and agent perspective-taking – as either (M_1) independent variable, (M_2) mediator variable, or (M_3) dependent variable as demonstrated in Table 12. The structural hypothesis predicts that an emphasis on actions *causes* moralization of the binding foundations and so we predict that M_1 will yield the strongest mediation model.

Table 12. *Candidate mediation models for comparison*.

	IV		*MV*		*DV*
M_1:	Agent focus	→	Binding concerns	→	Social conservatism
M_2:	Binding concerns	→	Agent focus	→	Social conservatism
M_3:	Binding concerns	→	Social conservatism	→	Agent focus

[20] Harm $r = -.050$, $p = .3$; fairness $r = -.054$, $p = .3$; loyalty $r = .078$, $p > .10$; authority $r = .101$, $p < .04$; purity $r = .144$, $p < .005$. In multiple regression models, the significant effects of familial political orientation on authority and purity foundations were non-significant, $p > .2$, after controlling for the effects of own political orientation on social issues, $p < .001$. As noted, familial political orientation did predict own political views: social $r = .380$, $p < .0001$; fiscal $r = .400$, $p < .0001$.

In our preferred model, M_1, agent focus causes concern for the binding foundations, which in turn promotes social conservatism. A mediation analysis revealed that concern for the binding foundations mediated the relationship between action focus and social conservatism, z = 7.46, p = 8.42 x 10^{-14}, *prop. effect mediated* = 1.02. By contrast, the alternative models, with agent focus as mediator (M_2) and as dependent variable (M_3) were only marginally significant, M_2: z = -1.77, p = .077, *prop. effect mediated* = -.067, M_3: z = -1.77, p = .076, *prop. effect mediated* = -.090.

We confirmed this relationship with our measure of evaluative simulation. Consistent with M_1, concern for binding foundations fully mediated the effect of agent perspective-taking on social conservatism, z = 4.77, p = 1.81 x 10^{-6}, *prop. effect mediated* = .854, whereas the alternative mediation models did not reach significance, z = .320, p = *.7, prop. effect mediated* = .008, M_3: z = .373, p = .6, *prop. effect mediated* = .054. Altogether, these analyses favored our hypothesized mediation model, M_1.

Given the ordering of the causal chain supported by the previous comparison, in the following step we present a path analysis with action and outcome foci as exogenous variables (*Level 1*), concern for the individualizing and binding moral foundations as intermediate variables (*Level 2*), and social political orientation as the dependent variable (*Level 3*). We seek to compare the variance explained by the full model retaining every path from a lower-level to a higher- level variable to the variance explained by a reduced model, where we drop the path from outcome focus to moralization of binding concerns, as well as the direct paths from evaluative foci to political orientation (see Figure 19). So, we computed the variance accounted for by each model using the following equation, $R_M^2 = 1 - \prod\left(1 - R_{P_i}^2\right)$, where $R_{P_i}^2$ is the R^2 from each regression in the model. The total R^2 of the full model equaled .397, while the total R^2 of the reduced model equaled .391, indicating a relative fit of the reduced model to the full model of .990. Model-fit difference was calculated using

$W = -(N-d)\times \log_e \left(1 - \mathrm{R}^2_{\mathrm{full}}/1 - \mathrm{R}^2_{\mathrm{reduced}}\right) = 4.720$, where N is the sample size and d is the number of dropped paths. We then obtained the $W_{crit} = \mathrm{X}^2_{crit}$, where $df = d$. For this analysis, $\mathrm{X}^2 = (df = 3, p = .05) = 7.815$. Since $W < W_{crit}$, we conclude that (i) the deleted paths did not contribute to the model and (ii) the reduced model fits the data as well as the full model.

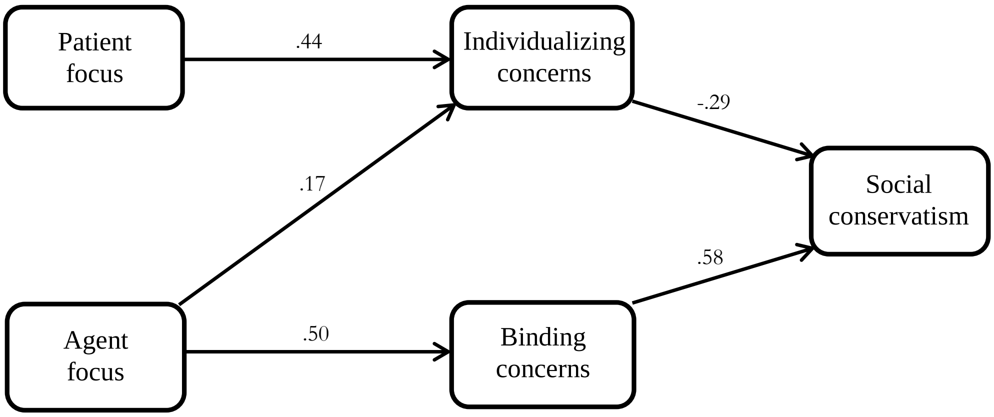


Figure 19. Reduced model showing β values as path coefficients.

5.6.2. Discussion

According to the above analyses, we propose the following account: Differences in agent and patient foci (associated partly with differences in cognitive style) give rise to distinct moral views about the individualizing and binding foundations. In particular, patient focus supports exclusively the moralization of harm and fairness foundations, whereas agent focus supports primarily the moralization of the binding foundations (and, to a lesser extent, the moralization of the individualizing foundations as well). In turn, moralization of the individualizing and binding foundations influences political affiliation: Specifically, individualizing concerns promote affiliation with the political left while binding concerns promote affiliation with the political right.

5.7. General discussion

Moral disagreements along the political spectrum appear to derive not only from a concern for divergent moral domains, but also from different structural foci during moral evaluation. Consistent with other recent findings (Graham et al., in prep; Piazza & Sousa, 2013), we show that conservatives are more likely than liberals to condemn direct harmful action when it is motivated by welfare-maximizing concerns, an effect not easily captured by moral foundations theory alone. A conservative preference for action-based moral evaluation is also evident in their agreement with abstracted statements (e.g., “By and large, morality is about doing what feels right”), and is associated with an approach to moral judgment that we call “evaluative simulation”: If it feels aversive to *me*, it’s morally wrong for *you*.

Additionally, we show that these evaluative foci in moral judgment are related to differences in thinking styles. In particular, the tendency to employ action-based affect in self-regulation may derive partially from having an intuitive cognitive style and lesser tolerance for uncertainty. This was true even after controlling for differences in political orientation.

Finally, we show that these structural differences between conservative and liberal approaches to moral judgment correspond with previously reported differences in the content of their moral values (Graham et al., 2009). Across different stimuli, we found that an agent focus was associated with moralization of the individualizing and binding moral foundations alike, whereas patient focus was associated exclusively with a concern for individualizing moral foundations. A focus on the intrinsic wrongness of actions is associated with valuing the binding foundations (ingroup loyalty, authority, and purity), while a focus on outcomes is associated with valuing the individualizing foundations (harm and fairness), and these relationships held even when controlling for political orientation.

At a broad level, these considerations highlight the importance of characterizing individual differences in moral judgment not just in terms of *domain content*, as in moral foundations theory, but also in terms of *evaluative focus*. Of course these two approaches are not mutually-exclusive. Rather, they complement each other insofar as certain contents are closely aligned with an emphasis on negative outcomes (particularly harm) and others are difficult to understand except in terms of an emphasis on aversive actions (most obviously purity).

The present study contributes to an extensive research program uncovering basic psychological differences between liberals and conservatives, e.g., in cognitive style (Kruglanski, 2005; Webster & Kruglanski, 1994) or in personality traits like threat anxiety, openness to experience, ambiguity tolerance (see Jost et al., 2001), and their longitudinal influence on the development of a political identity (Eidelman et al., 2012; Heaven et al., 2011; Hodson & Busseri, 2012; Matthews et al., 2009; Perry & Sibley, 2012). Building on prior evidence, we proposed the following causal picture: Reflective individuals tend to make moral evaluations by focusing on the expected outcomes of the agent's action on others, i.e., by evaluating the way that an agent's behavior affects the welfare of victims or beneficiaries. This approach to moral evaluation in turn gives rise to moral views, such as greater approval of welfare trade-offs and the moralization of individualizing but not binding concerns, that are sensitive to whether the behavior enhances or hinders patients' welfare. The development of this cluster of moral attitudes in turn promotes the individual's placement on the liberal end of the political spectrum. By contrast, individuals with a preference for intuitive and simpler thinking styles are more likely to make moral evaluations by focusing on the agent's action, i.e., by condemning the agent's behavior if its mental simulation elicits an aversive response. This approach to moral evaluation in turn gives rise to moral values that depend on whether performance of the agent's action would be aversive to the evaluator, such as the condemnation of welfare trade-offs and the moralization of binding concerns. The adoption of these moral attitudes then promotes the individual's identification as a social conservative.

Together these results might be considered to paint an unflattering portrait of the conservative moral sense. In place of a consideration of how people's actions harm or help those affected, conservative judgments about what is right and wrong for others depend on projecting personal, automatic aversions associated to others' behaviors: if it feels bad to me, it's wrong for you. At the same time, there are several potential benefits to an action focus in moral decision-making. First, ascribing intrinsic moral value to types of actions likely enables the adoption of bright-line standards of conduct, which are in turn associated with a greater ease of willful behavioral control (Baumeister & Tierney, 2011; Bennis, Medin, & Bartels, 2010). By contrast, evaluating the expected outcomes of an action introduces uncertainty and, potentially, a greater scope for processes of motivated reasoning to rationalize self-interested behavior. Second, evaluating others' actions by comparing them against the standard of one's own conscience would, at least in theory, reduce moral hypocrisy. Along similar lines, it might be expected to foster the development of morally and ideologically defined communities, a consequence that is compatible with the theory of conservative ideology as binding moral groups (Haidt & Graham, 2009).

We must also note several limitations of the studies presented here. First, the causal claims that we defend, while congruent with prior evidence and with the statistical models we tested, depend on correlational methods. Future work should therefore aim to test the proposed causal claims in an experimental design. Second, there are interesting public policy cases of moral disagreement where the structural approach makes the opposite (wrong) prediction. For instance, a paradigmatic emphasis on outcomes perhaps would yield the view that in certain circumstances the death penalty is morally preferable to life imprisonment, or that intrusions on privacy for the greater good of society (such as the practice of wire-tapping in the effort to prevent terrorism) are morally permissible. Yet the typically liberal position on these issues is in fact ordinarily the opposite. These examples illustrate that not all typically liberal moral views are congruent with those that a predominant outcome focus would promote. Finally, we must note that the effect sizes reported for content-based theories of conservative moral values tend to be larger than the effect sizes we report here based on structure. Of course, as we have emphasized throughout, these approaches are not at all exclusive of each other, and indeed they appear to be mutually-reinforcing.

Still, this study offers a promising advance in our understanding of the psychological basis of moral attitudes along the political spectrum. Liberals' and conservatives' seemingly irreconcilable views about matters of right and wrong are partly the product of individual differences in the tendency to approach the evaluation of moral issues with an emphasis on simulating the agent's action versus on assessing its expected outcomes. This difference in their approaches to moral judgment dovetails with other well-known correlates of political orientation, and helps to explain disagreements on heated public policy debates, hypothetical dilemmas, and abstract moral standards.

Chapter 6. The theory of evaluative focus in perspective.

6.1. Introduction

Throughout the previous chapters, I defended – through both argumentation and empirical evidence – that moral judgments involve affective responses to processes of mental simulation. The experimental findings presented in Chapters 4 and 5 demonstrate the utility of drawing a distinction between two particular kinds of mental simulation, differentiated by their content. Specifically, moral judgments may derive from a simulation focused on the agent's action or on the outcomes to patients.

To recapitulate, the studies in Chapter 4 showed that the condemnation of canonically disgusting and harmful behavior depends principally on the evaluative simulation of the agent's action. Then, in Chapter 5 we saw that differences in the evaluative focus on actions versus outcomes affects the moral (and even political) attitudes one adopts in fundamental ways. In particular, we saw that reason yields a preference for outcome-based assessments over action-based assessments, and therefore the endorsement of characteristically consequentialist and welfarist views, including the demoralization of loyalty, authority, and purity-related issues. As such, the theory of evaluative foci brings together two fruitful bodies of experimental research concerning, on the one hand, the interplay of automatic and controlled cognition in judgments about harm violations (Cushman et al., 2006, 2011; Greene et al., 2001, 2004; Paxton et al., 2012) and, on the other, the moral differences between liberals and conservatives (Graham et al., 2009; Inbar et al., 2009a). I will argue that this set of results is best construed as supporting a dual-process theory of moral judgment, implicating a System 1 responsible for the evaluative simulation of actions and a System 2 performing an assessment of its expected outcomes to victims and beneficiaries. On the proposed theory, the differences in moral attitudes observed in Chapters 4 and 5 are the result of differences in the reliance on and interplay between these neurocognitive systems (see Greene, 2007).

In order to do so, in this final chapter, I will integrate these findings in the context of a broader literature, ranging from cognitive neuroscience, through the literature on learning algorithms, and evolutionary theory. Through this effort, my aim will be to put forth a novel account of the psychological faculty for making moral judgments, and examine its plausibility and consistency in light of scientific theorizing about social cognition and decision-making across numerous, related disciplines.

6.1.1. Action and outcome as dual processes

As noted above, the findings in Chapters 4 and 5 provide empirical support for a fundamental distinction between two kinds of moral judgment. In particular, the observed differences in evaluative focus (which were associated with specific moral and political profiles) can be understood as reflecting underlying differences in the engagement and interplay between two neurocognitive systems.

The bulk of this dissertation has focused on deriving, and subsequently testing *behavioral-psychological* predictions about the human moral faculty. Yet, in Chapter 2, I outlined some of the broader postulates of dual process theories. A dual process interpretation of our findings makes certain assumptions also about the neural, evolutionary, and computational characteristics of each cognitive system. Therefore, the objective of this final chapter is to ask whether the proposed dual process theory is consistent evidence across this broader range of disciplines. For instance, do we observe dissociable and interpretable neural networks corresponding to the proposed cognitive systems for moral judgment? If so, does our characterization of System 1 as engaging an evaluative simulation of the agent’s action have a plausible evolutionary explanation? In this final chapter, I will examine a wealth of evidence from bordering disciplines – including neuroscientific data, computational models of learning and evolutionary considerations – to evaluate the broader tractability of this theory, and its consistency with findings concerning human cognition more broadly.

As part of this theoretical development, a central aim in this final chapter will be to explain an intriguing, and widely replicated, feature of the empirical data concerning moral judgment. As we saw earlier, in both original data I presented and numerous past studies, reason apparently privileges welfarist (and often specifically consequentialist) moral views. Participants who engaged critical thinking tended to discard concerns about loyalty, authority and purity (Paxton et al., 2012), and moralize concerns about harm and fairness only. Similarly, previous studies show that a tendency to engage rational thought is associated with permissive responding on trolley-type moral dilemmas (Feltz & Cokely, 2008; Moore et al., 2008). It appears therefore to be an empirical fact that rational individuals, as well as conditions that favor rational thought, converge towards consequentialist moral thinking.

The division between evaluative focus on actions and outcomes provides an explanation for this finding. Consider first the influence of reasoning on the adoption of moral foundations. We observed that more reflective participants tended to demoralize questions concerning loyalty, authority and purity, but not issues of harm and fairness, by comparison to intuitive participants.[21] According to the theory of evaluative focus, making moral judgments on the basis of an assessment of outcomes is favored by rational individuals for at least two related reasons. The first is that outcome assessment is more *cognitively demanding*, because for any single action there are numerous, foreseeable outcomes, each with their associated probability. The second is that actions

[21] Although we did not report the pairwise correlations between measures of cognitive style and moral foundations, we observe the expected pattern of correlations: Need for cognition correlated with demoralization of fairness, authority and loyalty, all *r*s > .3, all *p*s < .001. Meanwhile, need for closure correlated with moralization of fairness, authority and loyalty, all *r*s > .2, all *p*s < .01. Neither measure was related to participants' judgments about harm and fairness.

are temporally prior to outcomes, and some outcomes may be much delayed and even spatially distant. Therefore, participants who are likely to require minimal evidence in order to derive their judgment will use the first, available evidence while more reflective participants may be willing to await further information about outcomes. If we assess the morality of binding foundation issues – like flag desecration or homosexual sex – from an outcome perspective, we may not find any objectionable outcomes on which to condemn the behaviors. If instead we assess them by consulting our aversions to the performance of the evaluated actions, we are likely to find them more morally reprehensible. Therefore the tendency to engage in reasoning favors welfarist views, by supporting an evaluative focus on outcomes rather than on actions.

A recent study examining implicit attitudes toward homosexuality illustrates this account. Inbar and colleagues (2009b) found that people generally exhibit an implicit aversion towards homosexuality. (This was true even in individuals who profess tolerance on this matter, and of individuals who report a homosexual orientation.) When it comes to making *explicit* moral judgments, the typically conservative opposition to homosexuality is consistent with this System 1 aversion, and the theory of evaluative focus provides a plausible explanation for this result: A spontaneous, evaluative simulation of the action is responsible for the unconscious aversion to homosexuality, and conservative moral views are widely informed by this cognitive process. By contrast, in liberal individuals, the explicit acceptance of homosexuality is inconsistent with their implicit aversion, suggesting that some exercise of cognitive control is involved in professing a permissive attitude toward homosexuality.

In a similar fashion, the theory of evaluative focus explains why controlled cognition tends to promote utilitarian responding to moral dilemmas. In an outcome-based assessment, the outcome of five lives being saved (with certainty) is weighed against the outcome of one life being lost (with certainty), yielding the characteristically utilitarian judgment. By contrast, in an action simulation, the aversion associated to the harmful action of forcefully harming a person yields the characteristically deontological judgment (Miller et al., 2013).

So, reason privileges welfarist, characteristically utilitarian moral views, because reasoning and reflection yield the tendency to make moral judgments through outcome assessment. A central objective of this chapter will be to understand this empirical fact in greater depth: Why is reason associated with an assessment of outcomes rather than the simulation of actions? In the following sections, I will aim to answer this question by looking at several scientific findings about moral judgment from different domains.

6.2. Neuropsychology

As we saw in Chapter 2, one of the key premises of dual process theories is the existence and functioning of two neurocognitive systems in parallel (i.e., parallel processing). A few neuroimaging studies point towards the double dissociation of neural networks consistent with the theory of evaluative foci.

On the one hand, a neural network of regions generally engaged during System 1 processes is involved in making action-based moral evaluations (see Figure 20). For example, the left insula is activated during the experience of disgust (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003). In addition, the amygdala exhibits increased activation during the condemnation of purity-related issues (Schaich-Borg, Lieberman, & Kiehl, 2008). Similarly, the mPFC is activated during the condemnation of personal harm (Greene et al., 2001, 2004) and lesions in the mPFC decrease condemnation of personal harm (Koenigs et al., 2007). In addition, since conservatives demonstrate greater action focus, the theory of evaluative focus predicts that conservatives will exhibit a more robust System 1 network. This is indeed what a recent neuroscientific study finds (Kanai, Feilden, Firth, & Rees, 2011), with conservatives exhibiting more grey matter in amygdala and left insula than liberals.

On the other, a neural network of regions implicated in System 2 processes seems to be involved in making outcome-based assessments. This System 2 network is recruited in condoning utilitarian trade-offs in personal dilemma contexts (Greene et al., 2004; Shenhav & Greene, 2010). In addition, as predicted, liberals (who typically demonstrate greater outcome focus) exhibit also more gray matter volume (Kanai et al., 2011), as well greater connectivity (Amodio et al., 2007) in the anterior cingulate cortex than do conservatives.

So, altogether, the model of evaluative focus appears to be consistent with the relevant neuroscientific studies, integrating the evidence on neural activity during moral judgment tasks with functional and structural differences in the brains of liberals and conservatives.

6.2.1. Neuroscience of psychopathy

The study of the behavioral and neural correlates of psychopathy provides another valuable opportunity to test the model of evaluative foci. It is known that psychopathy is associated with greater endorsement of utilitarian trade-offs relative to non-psychopaths, particularly in *personal* trolley-type dilemmas (Bartels & Pizarro, 2011). We should therefore expect psychopaths to exhibit corresponding neural differences.

One possibility is that psychopaths are *hyper-reasoning*; i.e., they exhibit unusually heightened activity in the System 2 neural network that produces outcome-based assessments. Meanwhile, another possibility is that psychopaths have an *intuition deficit*: i.e., a structural dysfunction or decreased functional activity in the System 1, responsible for orchestrating action-based moral judgments.

The neuroscience of psychopathy draws support for the intuition deficit hypothesis (reviewed in Blair, 2008). In one study, psychopathic patients demonstrated reduced grey matter in regions of the prefrontal cortex, including the ventromedial areas of interest (Raine, Lencz, Bihrle, LaCasse, & Colletti, 2000). Meanwhile, other studies have confirmed amygdala dysfunction also in the pathology of psychopathy. In one study, high indices of psychopathy among violent offenders were associated with reduced amygdala size (Yang et al., 2009; see also Tiihonen et al., 2000). Functional deficiencies have also been documented: Psychopaths showed impaired amygdala activation during aversive conditioning (Veit et al., 2002) and emotional memory tasks (Kiehl et al., 2001). So, deficits in the System 1 network, which supports action-based moral judgments, are plausibly responsible for the tendency for psychopaths to condone personal harm in trolley-type contexts.

Together the variety of neuroscientific studies concerning moral judgment, political orientation and psychopathy indicate that the tendency toward utilitarian responding can derive from *either* a deficit in System 1 processes – as expressed by psychopaths (by comparison to psychotypical individuals) – or in increased System 2 activity – as expressed by political liberals (by comparison to political conservatives), echoing armchair accounts of the phenomenology of utilitarian thinking (Baron, 2011; Singer, 1981). At a broader level, these studies demonstrate the operation of two dissociable neural systems in moral judgment, compatible with the theory of evaluative focus.

6.3. Developmental acquisition of aversions to actions

The above findings document a deficit in System 1 processing in psychopathy, and suggest that the phenomenon of psychopathy may imply an impairment in action aversion, i.e., the absence of an aversion to harming others. A leading account of psychopathy, positing a *violence inhibition mechanism* (VIM), argues just this.

As I mentioned earlier, Blair and colleagues (1995) provide an explanation for how aversions to harmful actions are normally acquired throughout development. First of all, it is well known that psychotypical children react aversively to the presence of cues of submission or distress from an early age. I referred to this response as empathic concern (a kind of outcome aversion) in previous chapters. Throughout development, the psychotypical child may engage in aggressive actions which result in another's display of distress cues. These distress signals negatively reinforce the child's behavior, we might say, as *unconditioned stimuli*. But, through associative learning, the aggressive action may become negatively reinforced, as *conditioned stimulus*. So, consideration of the harmful action triggers an aversive, withdrawal response, which inhibits the action and decreases the likelihood of performing that harmful action in the future (Blair, 1995; see Figure 20).

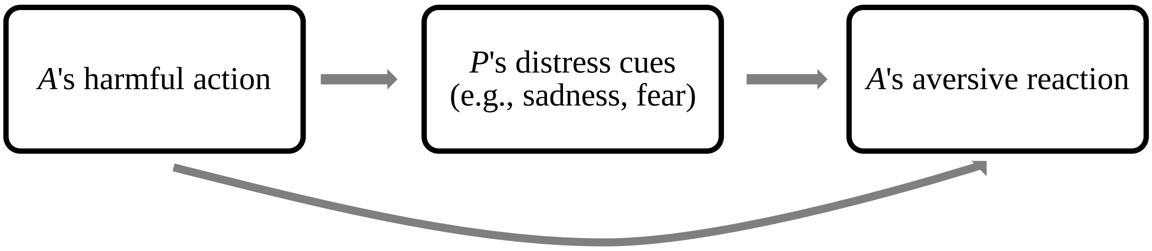


Figure 20. Schematic of the VIM demonstrating outcome aversion (i.e., empathic concern) and action aversion pathways.

This account suggests that deficits in the System 1 network should be expressed in impairments on aversive conditioning and instrumental learning tasks. Indeed, it appears that the effects of amygdala dysfunction include impairment in aversive conditioning and instrumental learning (LeDoux, 1998). Similarly, damage to regions of the prefrontal cortex results in difficulties learning associations (Bechara et al., 1994;

see also Blair, 2004). So the amygdala and prefrontal cortex putatively constitute the neural basis of the human system for associative learning (Blair, 2004; Damasio, 1999). On this account, the moral judgment patterns of psychopaths (Bartels & Pizarro, 2011) and patients with damage to the prefrontal cortex (Koenigs et al., 2007) are attributable to developmental difficulties in acquiring and storing the negative value associated to harmful actions.

This kind of explanation might account for the acquisition of aversions to commonplace violent behaviors, such as punching or kicking. Yet, there are at least two reasons why this mechanism cannot explain the acquisition of *all* aversions to actions. First of all, it appears that individuals have numerous moral aversions to actions that they have never performed themselves, such as shooting a gun (see Cushman et al., 2012). We might be able to save this explanation by appealing to visual representations and mental simulation. That is, we might acquire aversions, not only through own performance, but also through visual exposure to action-outcome pairings. For example, the repeated association of gun-firing to pain and death on television might plausibly, through aversive conditioning, result in the acquisition of an intrinsic aversion to shooting guns.

6.3.1. Beyond harm: aversions to victimless transgressions

Though this model seems well equipped to explain the acquisition of aversions to transgressions in the harm and fairness domains, there is reason to doubt its explanatory power with respect to aversions to victimless violations – for instance, to consensual incest, or to eating a pubic hair sandwich. Consider, for instance, the condemnation of flag desecration, or eating one’s dead pet dog. For these sorts of food- and sex-related transgressions there is no unconditioned distress response to the transgression (because there is no victim), and therefore no basis on which an aversion to the action itself may be conditioned.

Following prior theorists in the literature, in the following section (6.4) I will argue that some aversions to actions arise principally as a result of evolutionary pressures, rather than learned through conditioning. Transgressions in the purity domain, involving sex or eating, provide a particularly plausible case since the inhibition of these behaviors should confer some adaptive advantage to the agent him/herself.

At a broader level, in this chapter I will make the case that no single framework accounts for the acquisition of *all* moral aversions. Instead, moral aversions may derive from different concerns (some other-oriented, some self-oriented; see Dungan, Chakroff & Young, in prep.), and originate in different ways (principally through social learning, or principally through innate predisposition). This echoes a broader criticism of Elliot Turiel's (1983) work and of other developmental psychology perspectives. Crosscultural and evolutionary perspectives contravene Turiel showing that, at its root, morality is not comprised only of concerns about others' welfare (Haidt, Graham & Nosek, 2009; Kelly, 2013). One way of understanding the existence of moral concerns not derived ultimately from concerns about others' welfare, is to posit that some moral rules uphold *self*-oriented concerns (Dungan et al., in prep.). For instance, as evolutionary psychologists point out, one's goal of *avoiding contamination* might undergird numerous purity-related concerns. In addition, other self-oriented values such as the avoidance of punishment and the *maintenance of a positive reputation* in the moral community may undergird some of our moral norms as well. This might be particularly true about loyalty- and authority-related concerns, since violations of these domains have been shown to elicit others' contempt (Rozin et al., 1999b). So, just as victim distress plays an important role in the development of aversions to harmful actions, societal contempt may play a comparable causal role in the acquisition of aversions to loyalty and authority violations (such as, e.g., flag desecration).

Therefore, in order to retain the conceptual framework put forth by Blair to explain the development of action-based aversions across multiple domains, I suggest that there are numerous primitive (i.e., unconditioned) moral values from which action associations (or moral rules) are derived. The relevant unconditioned moral values include not only others' welfare but also self-oriented values such as the maintenance of a positive reputation. Furthermore, some moral aversions may not be the result of associative *learning* at all. Instead, some moral aversions may have an evolutionary basis, manifesting as the predisposition to acquire certain action-based aversions in the absence of sufficient social learning by association.

6.4. Innate predisposition: the case of incest

Several leading psychologists, anthropologists, and primatologists have held versions of this so-called *adaptationist* view in evolutionary approaches to morality (de Waal, 1996; Katz, 2000; Sober & Wilson, 1999). On this view, selection pressures which were present in evolutionarily ancient environments have likely predisposed us to acquire aversions that inhibit putatively maladaptive action-types. This perspective is not without its detractors (Prinz, 2007), but in what follows I will try to draw out its plausibility, if limited to certain purity-related issues (Rozin, Haidt, & McCauley, 2008). To do this, I will examine the moralization of incest as a case study; although I think one can extrapolate this account to several other food and sex-related taboos. Three features of the moralization of incest point toward an innate predisposition: (1) the *norm independence* and (2) *adaptive advantage* of incest avoidance, as well as (3) the *psychophysiology* of disgust.

Incestuous behavior is widely condemned, and in many cases also legally sanctioned, in human societies worldwide (Wolf & Durham, 2004). Of course the near-universality of the incest taboo does not entail that it is innately predisposed. More compelling evidence derives from the apparent *norm independence* of incest avoidance. As we saw in Chapter 3, even in sociocultural environments such as the Israeli kibbutzim, where no explicit norm governs incestuous behavior, it is strikingly rare (Shepher, 1983). Not only do kibbutz-raised children experience sexual revulsion towards their siblings, but also towards other non-siblings with whom they were communally raised. This finding provides support for the hypothesis, first proposed by Westermarck (1891), according to which incest avoidance relies on a mechanism for sibling detection that gauges genetic relatedness as a function of the duration of coresidence during childhood. Additionally, this mechanism might explain the unusually high divorce rates among Taiwanese minor marriages, which are characterized by an extended period of premarital cohabitation between the future bride and groom (Wolf, 1995). In sum, incest avoidance appears to arise in the absence of a social norm, driven by a sibling detection mechanism.

The *adaptive advantage* of incest avoidance and condemnation also suggests an innate basis for this behavioral tendency. Inbred offspring exhibit reduced genetic variance, and thereby are more likely to suffer reduced fertility, loss of immune system function, increased infant mortality, and lower general fitness. We observe this clearly in enclosed habitats, such as islands, where animal communities are forced to inbreed. As this process goes on over generations, populations tend to decrease dramatically and interbreeding between remaining individuals becomes harder to avoid, leading to a relatively inevitable extinction (Frankham, 1998). Therefore, from an evolutionary perspective, an automatic and perhaps over-inclusive mechanism of incest avoidance confers an adaptive advantage. In line with this, several mammal species avoid inbreeding, particularly when non-related partners are available (Wolf & Durham, 2004).

Adaptations in behavior are often marked by an automatic, psychophysiological reaction that initiates the appropriate behavioral response, e.g., of approach or of avoidance in relation to the elicitor. This is evidently the case with the avoidance of cliff edges (Dahl et al., 2013), spiders, or snakes (LoBue & DeLoache, 2008; Ohman & Mineka, 2001). These adaptations are marked by facial and cardiorespiratory routines that initiate the appropriate states of alertness and avoidance in response to threats that would have been present in ancient environments. Similarly, incest (as an example of a purity-related transgression) has an associated psychophysiological response, which we know as disgust, and observe clearly in laboratory settings (Haidt et al., 2000; Rozin et al., 1999b). The disgust response can be traced to its psychophysiological precursor, *distaste*, which serves principally against oral contamination and is widely observed in other species (Rozin et al., 2008). Disgust arises automatically in response to contaminant stimuli and also maladaptive practices – incest, zoophilia – and features a particular facial expression, as well as some behavioral responses (including vomiting), which serve to reduce chances of contagion. By the very nature of contagion, if a community member is contaminated it may not be long before I am contaminated too. Therefore, it does not much matter whether it is I or some community member who is contaminated and so the disgust response is triggered comparably by others' transgressions as it is by one's own. This feature renders disgust a patently *social* emotion.

Altogether, considerations of the norm independence, adaptive advantage, and psychophysiological profile of incest avoidance suggest that the mechanism is the result of selection pressures. This favors an adaptationist explanation, over a social learning, explanation of the acquisition of an action aversion towards incestuous behavior. A similar case could perhaps be made about other food- and sex-related taboos that share the above characteristics. Moreover, the adaptationist explanation may even hold some ground outside the purity domain (de Waal, 1996). For instance, Trivers (1971) famously argued that "tit for tat" policies provide an adaptive advantage, and so related views of fairness may have an evolutionary basis. Similarly, Greene (2007) has argued that the inhibition of intentional and forceful harm may have an innate basis. I remain agnostic about the strength of adaptationist perspectives outside the purity domain; my goal in this section was merely to demonstrate that at least *some* moral attitudes are predicated on action aversions that we are innately predisposed to acquire.

6.5. Aversions in practice: Behavioral choice and reinforcement learning models

Throughout previous sections, I have argued that aversions to actions are acquired in a variety of ways. Following Blair's violence inhibition mechanism, some aversions – particularly those belonging to the harm domain – are conditioned from the basic empathic concern for distress in conspecifics. Other aversions to action-types are conditioned from self-oriented concerns, such as the avoidance of societal contempt and the maintenance of a community reputation. And, others yet, most obviously in the purity domain, may be the product of evolutionary pressures. These aversions are acquired through numerous distinct pathways and result in the grab bag of action aversions typically expressed by normal adults.

The literature on machine learning has identified two broad approaches to behavioral choice, known as model-free and model-based algorithms (Sutton & Barto, 1999; see also Sutton, 1988). In this section, I will argue that these reinforcement learning algorithms provide a plausible framework to understand the engagement of action and outcome aversions in moral judgment across different domains. This broader framework enables us to understand how action and outcome aversions across domains might be brought online during behavioral choice (see also Crockett, 2013; Cushman, 2013). Specifically, model-free and model-based algorithms engage action and outcome aversions correspondingly.

One class of algorithms, known as *model-based*, employs probabilistic models of the causal relations in the world to select behavioral choices (or sequences of behavioral choices in some cases) that lead to desired rewards (and away from punishment). Model-based algorithms for behavioral choice produce instructions with the following structure:

> Perform action A because a sequence of actions beginning with action A will lead over time to the maximal set of rewards, $\sum_{i=1}^{n} R_i$,

followed by a specification of the full temporal sequence of actions and their corresponding rewards and/or punishments: "Pushing the man will yield negative value and also cause the man to fall on the tracks, which will cause the train to stop, which will cause the five workers to be saved and yield larger positive value. Not pushing the man will cause the man not to fall on the tracks, which will allow the train to proceed, which will cause the five workers to die and yield large negative value. So, push the man off the footbridge."

This is essentially the kind of decision process that we have described as outcome-based evaluation in the moral domain. In moral judgment, outcome-based evaluation select actions based on the expected value of their corresponding outcomes, where these outcomes primarily concern the welfare of victims and beneficiaries (but perhaps other self-oriented moral values too, including one's own community reputation and welfare).

The alternative algorithm for behavioral choice, known as *model-free,* builds sparse representations of the value of each behavioral option available in a particular state. The basic structure of model-free algorithms is the following:

> In context *C*, among the actions available, perform action *A* because *A* is associated with the highest reward, R_A.
>
> In any state, the choice among available actions depends solely on the

representations tied directly to those specific actions in that particular state—without consulting the outcomes of the actions: "Pushing the man will yield a negative value. Not pushing the man has no associated value. So, don't push the man off the footbridge."

Rather than performing searches over the enormous space of possible future actions, model-free algorithms simply query the values of each of the behavioral options that are immediately available, much like the account of action-based evaluation I put forth throughout prior chapters. In moral judgment, action-based evaluation select actions based on the value associated intrinsically to the action itself, as a product of innate predispositions, or associative pairing from signs of victim distress or of societal contempt. This account therefore extends Blair's model of associative learning in the harm domain to explain the acquisition, whether through ontogeny or phylogeny, and engagement of action and outcome aversions across numerous moral domains.

6.5.1. Devaluation procedure

The operation of these two algorithms is neatly demonstrated by the *devaluation procedure*, an experimental paradigm that has been conducted with rats, monkeys and humans (Schultz, Dayan, & Montague, 1997). In a simple version of the devaluation procedure, Dickinson and colleagues (1995) trained a group of rats to press a lever in order to receive food. Half of the rats received *basic* training (consisting of 120 trials), while the other half received *extensive* training (consisting of 360 trials). Extensive training serves to strengthen the representation of the value of lever-pressing. This is because the more trials a rat undergoes, the more the rat will associate a reward (i.e., the food) to the action of pressing the lever and come to value lever-pressing in itself.

In each group, prior to the experiment, half of the rats were fed while the other half were not fed. The rats that were fed constitute the group for which the outcome was *devalued*, since the reward is no longer as valuable as it is for the unfed rats in the *valued* group. Researchers measured the number of times per minute that the rats in each group pressed the lever, as an index of the value assigned to the option of pressing the lever. To the extent that model-free algorithms determine behavioral choice we should expect that rats press the lever even if the reward has been devalued. Meanwhile, to the extent that model-based algorithms determine behavioral choice, we should expect the rats to press the value as a function of whether the reward is valued or not.

The researchers observed that the rats who had received basic training, and had relatively weaker value representation of lever-pressing, pressed the lever as a function of whether they valued the outcome, i.e., more frequently when they were hungry than when they were full. By contrast, the rats who had received extensive training and therefore had built up a strong value representation associated to lever-pressing intrinsically, tended to press the lever equally regardless of whether they had recently been fed. This suggests that, as intrinsic value representations of the action were strengthened, model-free algorithms began to drive behavioral choice.

6.5.2. Model-free versus model-based moral evaluation

As we saw exemplified by the devaluation procedure, model-based algorithms equipped with a detailed model of the causal world and unlimited processing capacity can make what we might call *rational* choices. This owes to at least three features of model-based algorithms; they are (i) far-sighted, (ii) goal-oriented, and (iii) flexible. They are *far-sighted* choices because they can specify a decision path that may require a long sequence of actions to reach the desired goal. They are also *goal-oriented*, in the sense that decision paths are fully determined by the values associated to end states (i.e., by the goals) held by the agent. Finally, they are *flexible*, in the sense that decision paths can be updated easily as new punishments and rewards are introduced to the model.

At the same time, there is a high *computational cost* associated with constructing and operating a causal model of a complex world. As the number of relevant states and actions increases, so does the space of possible decision paths over which a model-based algorithm must search. This renders a model-based solution to numerous real-life decisions practically untenable. In these contexts, model-free algorithms, with their light computational load, tend to dominate behavioral choice.

However, by merely storing value associations related to each behavioral option, model-free decisions can often lead to sub-optimal outcomes. We saw this in the devaluation procedure, when the actual reward did not match the stored value associations. So how are these sparse value representations updated to yield beneficial choices in a changing environment? This is a serious challenge for model-free algorithms and computational models demonstrate two techniques that update model-free value representations in order to minimize error. *Prediction error learning* and *temporal difference learning* enable model-free algorithms to more efficiently represent the value of a behavioral choice, without representing what specific rewards make it so.

Prediction error learning (PEL) allows an agent to update a representation of the value of an action based on recent trials of that action-type. So, for example, suppose an agent has a stored value representation for a certain action of 10, and a new token of that action-type yields an actual reward of 4. PEL allows the agent to update the representation of the value of that action-type by some fraction of the error, e.g., to 8 (see Schultz et al., 1998).

The process of *moralization*, whereby moral attitudes towards an action-type shift dramatically and relatively quickly, reflects the operation of prediction error learning in moral judgment. Consider, for example, the morally motivated transition to a vegetarian diet. During this transition, vegetarians are likely to develop a disgust reaction to meat-eating, and this disgust response helps to conform to the vegetarian diet (Rozin, Markwith, & Stoess, 1997). The development of a categorical disgust response towards meat-eating can be understood as the result of PEL. As individuals acquire vegetarian moral values, they effectively introduce a new "punishment" for eating meat into their models of the world. When this punishment is newly introduced, the stored model-free representation of the value of meat-eating is likely greater than the value of its actual (i.e., newly devalued) outcome. So initially the avoidance of meat-based foods depends on a model-based computation of the harmful outcomes for animal welfare (or, in simple terms, on the active reminder of its consequences for animal suffering). But, through PEL, the vegetarian will update the model-free value representation of meat-eating, resulting over time in an aversion to the *action* of meat-eating over and above any aversion to animal suffering.

But, in the real world, rewards and punishments often result from a *sequence* of actions. For instance, receiving an acceptance letter from university leads to going to university and going to university leads to a number of rewards such as becoming educated, skilled, and/or employed. Through model-based reasoning one might be motivated to apply to university because this sequence of causal relations leads to the rewards we aim to obtain. But, since applying to university does not directly lead to any rewards, it cannot through associative pairing acquire value. So, how might model-free algorithms help us to decide whether to apply to university? *Temporal difference reinforcement learning* (TDRL; Sutton, 1988) enables a model-free algorithm to guide behavior towards rewards that require a sequence of two or more actions. TDRL assigns value to a behavioral option that does not *directly* yield a reward/punishment, but rather alters the probability of attaining a reward/punishment by leading to another behavioral option that yields the reward/punishment (as in the above case). So by treating intermediate actions as rewards, TDRL leads the agent through sequences of actions to distant rewards.

In the moral domain, TDRL appears to play some role too – for example, when we condemn the formation of malicious intent or desires. That is, (Option *B*) carrying out a plan to murder *S* leads with probability P_1 to *S*'s death, but (Option *A*) desiring or intending to murder S leads to (Option *B*) carrying out a plan to murder *S* with probability P_2. TDRL treats Option *B* as inherently morally wrong, and thereby enables us to ascribe moral wrongness to the Option *A*, rather than treating it as a mere antecedent to *B*.

Even while these techniques enable model-free algorithms to update their value representations in beneficial ways, and retain their computational frugality, it is clear that model-free algorithms remain error-prone by comparison to model-based evaluations. As effective as they are in maximizing rewards when environments are stable over time, they are much more likely to err when deployed in conditions that differ from those in which the association was established. That is, when these associations are brought online in novel environmental contexts, such as consensual sibling incest cases or trolley-type dilemma contexts, model-free approaches typically lead to notable prediction error. Here, only a model-based analysis of expected outcomes can yield an accurate judgment of the value of the target behavior.

6.6. Evaluative simulation in action

So far in this chapter, I have discussed the origin of our aversions to actions and outcomes, and the cognitive processes by which they are engaged in decision-making and behavioral choice. In this final section, I want to examine a question I have set aside up to this point concerning the role of aversions specifically on *moral* evaluation: Why do our personal aversions to actions ("It would be disturbing for me to have an incestuous relationship.”) shape *moral* judgments about others’ behavior (“It is wrong for Jenny and Rick to have an incestuous relationship.”), above and beyond influencing individual behavior and decision-making (“I would not have an incestuous relationship.”)?

First of all, it is worth pointing out that we rather ubiquitously simulate the behavior of others, and we do so often in order to understand them (Goldman, 2006). We spontaneously yawn when others yawn, cry when others cry, and other times we more deliberately “put ourselves in another person’s shoes”. Therefore, assuming that a certain moral transgression (when considered as an action plan for the self) produces an aversive response, and that we rather spontaneously simulate others particularly to understand their behavior, it is reasonable to suppose that another’s performance of said moral transgression should result in a comparable, aversive response in the self (as posited by evaluative simulation).

Second, evaluative simulation might arise rather naturally also due to the broader merits of model-free thought reviewed already. As we saw, model-free evaluation is less cognitively demanding than is model-based evaluation (generally). And this should presumably be true of action-based valuation in moral cognition, whether these processes contribute to the regulation of own behavior (“I should not push the fat man off the footbridge”) or the evaluation of third-party behavior (“It is morally wrong for Suzy to push the fat man off the footbridge”). On a related note, while the outcomes of a given action can be temporally and spatially distant from their corresponding causal agent, in normal cases actions are immediately visible and easily attributable to the causal agent. Therefore, just as we routinely regulate our own behavior by assessing aversions to performing actions (rather than on the value of its outcomes) *because it is cognitively ‘easier’*, we may tend to judge others’ behavior through a similar process (i.e., evaluative simulation) for the same reason.

These considerations point toward the *heuristic* value of evaluative simulation (see Kahneman & Frederick, 2002, 2005). This strategy enables us to determine whether other people's behavior is harmful or immoral – with remarkable ease and success – without having to consider the numerous outcomes that result from it. But, the above considerations do not yet explain why individuals would be motivated to morally judge others in the first place.

This specific discussion quite naturally adopts an evolutionary perspective. As we saw early in Chapter 2, it is a common feature of System 1 cognitive processes that they arose relatively early in our evolutionary history, and can be understood by pointing to one or more adaptive advantages they are thought to have bestowed individuals in ancestral environments.

So, the presumption that a certain psychological faculty is automatic and unconscious prompts speculation about its evolutionary origin and function. In these final paragraphs of my dissertation, I will set aside the emphasis on providing a *mechanistic* account of evaluative focus, to instead comment on its putative *functional* role. In other words, what adaptive advantages might have promoted the selection of evaluative simulation as a strategy for third-party moral evaluation? Why – from the functional perspective, once again – might people judge what is right or wrong for *others* on the basis of what feels right or wrong for *themselves*? So, in closing, I will make some brief and indeed speculative remarks on the evolutionary basis of evaluative simulation. I will suggest a few advantages that evaluative simulation would have bestowed on individuals in early, social communities, promoting its selection as a psychological adaptation. Specifically, I will argue that evaluative simulation (1) promotes conformity to group norms, (2) reduces risk of infection from external pathogens, and (3) boosts interpersonal trust. The purpose of this exercise is not to convince the reader of the details of my suggestions, but rather to chart new terrain at the intersection of morality and evolution, putting forth a series of related hypotheses for which the evidence, though favorable, provides only indirect support.

6.6.1. Incentivizing conformity, fostering homogeneity

It is clear that upholding moral standards can demand a great effort, especially in circumstances that pit moral values against other prudential interests. Behaving in congruence with one's moral views frequently demands self-control and the exertion of willpower, as Aristotle famously defended. Contemporary psychological research vindicates this perspective, showing that willpower is a limited resource which is consumed in circumstances that require self-control (Baumeister et al., 1998).

In this sense, the judgment of others – both of blame and of praise – might provide an added incentive to behave normatively. Indeed, it is patently clear that people modify their behavior in the presence of others, often in order to exhibit more socially desirable traits. Therefore, the projection of internalized aversions onto others' introduces an added incentive (for them) to inhibit moral violations and, therefore, should have the effect of promoting conformity to prevailing moral standards in the community. In this way, a community of *moralizing* individuals should achieve greater conformity to its shared normative standards than a non-moralizing community, in which individuals are left to self-regulate their behavior.

We find suggestive evidence for this effect in United States culture, where tobacco smoking has recently been subjected to heavy moralization. Retrospective reports indicate that disgust directed at smoking has increased in recent times, along with scientific knowledge about its harmfulness (Rozin & Singh, 1999). However, Rozin and Singh (1999) found that it was the strength of one's disgust aversion, and not beliefs about the harmfulness of smoking, that most robustly predicted attitudes of moral condemnation towards smoking behavior. This finding suggests, as the authors argue, that the associative pairing of an action to the feeling of disgust leads to moralization of the action. Critically though, owing to this process of moralization, rates of smoking appear to have decreased dramatically in the United States (Saad, 2012), by comparison to other countries where smoking is not as notoriously moralized (Helweg-Larsen & Nielsen, 2009). We might think of this function as that of providing an *incentive to meet normative standards*.

The failure to conform comes at a high personal cost to the wrongdoer, since when we condemn someone's immoral behavior ("It was wrong for Jenna to steal Sally's packed lunch"), we often form parallel impressions about them and their character ("Jenna is selfish and vicious"; see Bar, Neta, & Linz, 2006; Funder, 2004; Todorov, Said, Engell, & Oosterhof, 2008). This characteristic arises remarkably early in life (Kuhlmeier, Wynn, & Bloom, 2003), and across a variety of cultures (Fiske, Cuddy, & Glick, 2007; Hamlin, Wynn, & Bloom, 2007). Moreover, these character evaluations apparently play a critical role in our decisions about who to interact with (Gintis, Henrich, Bowles, Boyd, & Fehr, 2008; Rand, Dreber, Ellingsen, Fudenberg & Nowak, 2009). This observation brings us to a distinct, but related, function of evaluative simulation: constituting one's moral ingroup. To see why, consider the following hypothetical scenario:

> A small-scale community lives on a remote island. Early in the history of this society, many people contracted a deadly virus and died as a result of eating the rather delicious meat of an abundant boar. The islanders soon ceased to eat the animal's meat, which restored normal mortality rates in the community.
>
> To avoid future risks to the community, children were told stories of the nasty animal – which became known as the vilebeast ("vile beast") – and how it was planted on the island by the Devil. Within a few generations, the islanders were avoiding the animal altogether. Even the sight of a vilebeast was disturbing to most. The premature death of an islander, particularly during periods of food scarcity, would often prompt rumors as to whether he or she had been "taken by the vilebeast". One day, a group of people are observed eating what most certainly is a vilebeast stew.

In Universe A, the islanders practice evaluative simulation, so they judge the group's behavior to be morally reprehensible. Unless the minority conforms to the group's prevailing norm, they will be ostracized from the community. In Universe B, the islanders practice outcome assessment, and so judgments of the group's behavior depend on outcome information that is not immediately accessible.

As noted before, evaluative simulation holds the advantage of determining group membership with marked ease. The mere performance of devious behavior is a sign of outgroup membership. As it were, in many cases, these devious actions *do* probabilistically result in negative outcomes, but in a minority of cases deviant behavior will turn out harmless (e.g., if the vilebeast eaters were immune to the virus, or had discovered how to examine their game for signs of the virus). So, whereas deviant behavior will be eradicated in Universe A, it will be tolerated to the extent that it is harmless in Universe B.

This conclusion can be (albeit simplistically) modeled by defining the homogeneity of the islander population as the probability that an ingroup member follows all moral rules. Suppose also that the islanders follow a one-strike policy: In Universe A the evaluative simulators will excise an individual after breaking a moral rule, while in Universe B outcome assessors will do so after the individual causes a morally bad outcome (by comparison to other outcomes he/she could have brought about by acting differently at that time). Once again, the homogeneity of the islander community in Universe A will invariably be 1, since no individual can break a moral rule and belong to the ingroup. Meanwhile, in Universe B, the homogeneity of the islander community will depend on the combined probability that following a moral rule leads to the morally preferable outcome, $\prod_i^n p_i$.

For simplicity's sake, let's suppose that the islanders have a total of twenty moral rules ("ye shall not eat vilebeast", "ye shall not curse the leader's name", "ye shall open your house to other islanders during the monsoon season", and so on), and that – for *every* moral rule – following the rule leads to the morally preferable outcome with a probability of .97, i.e., almost every time. This reduces the homogeneity in Universe B to almost half (.54). In other words, we would expect just under half of the population to engage in *at least one* form of harmless deviance! There are some flagrant simplifications in this exercise, but the generalization still holds: evaluative simulation brings about a remarkable degree of homogeneity (by comparison to outcome-based judgment).

Circumstantial evidence lends support to this idea. As we saw in Chapter 5, evaluative simulation is relatively more popular among political conservatives than among liberals. We should therefore observe that conservative individuals prefer more homogeneous moral communities than do liberal individuals. To examine this question, I re-analyzed a 2003 study on religious diversity from the Association of Religion Data Archives (ARDA, 2003), comparing the responses of 992 Republican voters (i.e., conservatives) to the responses of 785 Democratic and Green party voters (i.e., liberals). On all twenty items, Republican voters were more likely to value religious and cultural homogeneity than were Democratic and Green party voters (.21 < all Cohen's *d*s < .64; see Supplementary Analysis 4). For instance, they were more likely to object to their children marrying a Muslim "who had a good education and came from a good family", or to oppose the construction of a Hindu temple in their neighborhood. Similarly, conservative participants were more likely to believe that immigration should be limited, and that immigrants should be acculturated (i.e., "give up their foreign ways and learn to be like other Americans") and non-Christians converted to Christianity. At a broader level, Republican voters tended to hold the belief that religious diversity has not been "good for America", and that it is a threat to traditional values. A separate household survey confirms that conservatives are more likely than are liberals to prefer living in an area where most people share their political views, race, and religion (Pew Research, 2008).

Therefore, the collective exercise of evaluative simulation may give rise to relatively homogeneous moral communities. This was suggested by an initial thought experiment, and then tentatively supported by recent sociological surveys. In future work, the precise relationship between evaluative simulation and community homogeneity should be examined more directly however. Does racial, cultural and religious homogeneity arise in conservative-dominated areas with geographical *county* as the unit of analysis, for example? In addition, it is important to delve further into the mechanism by which the communal exercise of evaluative simulation yields homogeneous groups. I suggested that through evaluative simulation individuals will tend to value normative behavior and derivatively others who behave in congruence with their own norms. Conversely, they will punish even harmless deviance from normative standards by exclusion from the ingroup. Whether this model explains the proposed relationship between evaluative simulation and community homogeneity is a matter that awaits more systematic investigation.

6.6.2. The value of moral homogeneity

What is the adaptive advantage of community homogeneity, then, which would lead to the selection of evaluative simulation as a psychological adaptation? What is the value of homogeneity over heterogeneity in social networks (such that homogeneity might have been an adaptive feature of ancestral communities)?

In the *Phaedrus*, Plato famously noted that “similarity begets friendship”. Indeed, a number of recent proposals concerning the evolutionary origin of morality converge on the role of homogeneity in strengthening social networks. For instance, proponents of the *green beard hypothesis*, like Dawkins (1976) and Frank (1988), argue that the emergence of moral behavior might have depended on a genetic adaptation that yielded (a) a phenotypic signal of membership in the moral community, and (b) a tendency to behave altruistically towards carriers of the phenotypic signal. In other words, the success of moral comunities would require some sort of phenotypic homogeneity, such as a green beard, that signals the disposition toward prosocial behavior. Meanwhile, Johan Koeslag (1990, 1997) argued that *koinophilia* – or the widespread psychological adaptation leadings individuals to prefer interaction with others who display *normal* traits – helps to explain numerous evolutionary phenomena, potentially including the origin of human cooperation. Since mutations of a certain phenotypic or behavioral feature can co-exist, evolutionary pressures cause ecologically beneficial features to dominate while disadvantageous counterparts become increasingly rare. Therefore, predominance of a trait in a population is associated with adaptiveness just as rarity is associated with maladaptiveness. This phenomenon, Koeslag thought, might explain why we are unusually attracted to average faces (Langlois & Roggman, 1990; Apicella, Marlowe, Fowler, & Christakis, 2012), even from an early age (Rubenstein, Kalakanis, & Langlois, 1999), and why animals are speciated into phenotypic clusters (Koeslag, 1990).

These accounts converge on the role of homogeneity in the origin and success of ancestral, moral communities. In this section, I will point towards two principal virtues of homogeneous social networks, i.e., disease management and social trust, which plausibly derive from the collective exercise of evaluative simulation and might contributed to the success of ancestral moral communities.

Managing disease and contagion

Consider again the case of the islanders and their prohibition of vilebeast-eating. As we noted, in Universe B, (where outcome assessment is the predominant approach to moral evaluation) the islanders suspend their moral judgment of the deviants until the outcomes of the action are observed. This is evidently disadvantageous for the community, insofar as it introduces a risk of contagion to the community.

Therefore, one putative function of evaluative simulation is to prevent the introduction of contagious and maladaptive behavior to the community. Some evidence supports this: The prevalence of religiosity and authoritarianism, and the strength of family ties – all well-known correlates of social conservatism and evaluative simulation – are predicted by levels of pathogen stress, both in comparisons between different countries and between different states in the United States of America (Fincher & Thornhill, 2012; Murray, Schaller, & Suedfeld, 2013). Areas with high levels of parasite stress tend to be inhabited by more religious and authoritarian societies than are seen in areas with less parasite stress (even after controlling for potential confounds such as human freedom and economic development).

This implies that societies in high parasite stress regions exhibit greater emphasis on loyalty-, authority- and especially purity-related norms. So, binding concerns might arise as a response to greater infection risk in conditions of heightened parasite stress. This suggests a role for evaluative simulation in managing disease stress, in two principal ways: (1) by thwarting contagion within the ingroup through the condemnation of various purity-related transgressions, and (2) reducing infection risk from foreign pathogens by eschewing contact with dissimilar, outgroup members. In sum, where contagion might be a major concern, the kind of flat prohibition of deviant behavior facilitated by evaluative simulation is a better, i.e., more adaptive, approach to evaluating others.

Boosting social trust

Finally, another proposed advantage of community homogeneity is the resulting increase in social trust (see Hamlin, Mahajan, Liberman, & Wynn, 2013; Mitchell, Macrae, & Banaji, 2006). Sociological surveys comparing social diversity across several United States (Alesina & La Ferrara, 2002; Putnam, 2007) and cross-national (Delhey & Newton, 2005) locations provide some support for this supposition:

Diversity is associated with decreased public engagement and social trust, whereas homogeneity seemingly favors public engagement and social trust.

Koeslag (1997) provides a neat explanation for this, by considering the Iterated Prisoner's Dilemma (IPD). As noted earlier, according to the theory of koinophilia, individuals prefer others with predominant rather than rare characteristics – essentially because most mutations reduce fitness. In an IPD game, this preference has the effect of increasing the fitness of the common strategy, and stabilizing almost any strategy that has become the local norm.[22] In a similar sense, the green beard hypothesis (Dawkins, 1976; Frank, 1988) argues that the players' phenotypic expression of a green beard in an IPD game should yield social trust, essentially, because it enables the expectation of other players' cooperation.

Koenslag's (1997) explanation hinges on the possibility of group-level selection, i.e., that a behavior (in this case, the spontaneous projection of aversions) can spread in a population because of the benefits it bestows on the group (in this case, conformity to prevailing norms, disease management, and social trust) despite incurring a fitness cost at the level of the individual. Adaptationist views of morality, like the one I am defending, often make this assumption (Bowles & Gintis, 2011; Haidt, 2012; Henrich, 2004; Wilson & Wilson, 2008). Still, as we saw in the green beard view, community homogeneity might be expected to result even on a gene-level explanation. Therefore, a psychological adaptation that – like evaluative simulation –fosters homogeneity in the

[22] Different, partially isolated communities of individuals will therefore evolve different strategies (some more cooperative than others) and defend its strategy against alternatives that may arise through mutation or contact with other communities. E.g., selfish individuals will be ostracized from cooperative communities, because of their deviant and unusual behavior, and the advantage of exploiting their cooperative peers outweighed by the costs of ostracism, not finding a mate, and so on. Finally, by the very nature of cooperation, groups that happen to stabilize a cooperative strategy will be fitter than groups who have stabilized defection (Koeslag, 1997).

community could help to explain the evolution of cooperation and social trust, either at group-level or gene-level selection.

Still, much more work is required to test the various hypotheses I have outlined in this section. I hope only to have highlighted some future avenues of research at the intersection of morality and evolution. First and foremost, it is plausible that communities governed by evaluative simulation exhibit heightened homogeneity. In addition, several proposals converge on the hypothesis that community homogeneity played a role in the emergence of morality in ancestral communities, and I highlighted two possible reasons why: (1) improved disease management, and (2) heightened social trust. My argument hinged on a variety of evolutionary suppositions for which direct evidence is notoriously difficult to ascertain. So in this chapter I have sought to test these ideas indirectly through proxy evidence about religious and conservative communities. To the extent that religious and conservative individuals employ evaluative simulation, and construct homogeneous moral ingroups, they might reflect the mechanisms by which the collective exercise of evaluative simulation contributed to the emergence of ancestral, moral communities.

6.7. Conclusion

I will conclude by way of a final restatement of the theory I defend in light also of evidence from the neighboring disciplines discussed in this chapter. According to the dual process theory of evaluative foci, moral judgments are the product of the interplay between two neurocognitive systems. A System 2, marked by activation in the anterior cingulate cortex and dorsolateral sections of the prefrontal cortex, conducts a model-based assessment of the value of the outcomes derived from each action. This system depends on the assignment of value to certain outcomes that are deemed morally relevant, most notably, the welfare of other humans. Meanwhile, a System 1 –

implemented most clearly by a network that features the amygdala and ventromedial prefrontal cortex – conducts an evaluative simulation of moral behavior, engaging model-free representations of the value associated to the target actions. These model-free value representations implicated in action-based assessments might arise in a number of ways: some through associative learning from fundamental moral concerns such as distress, and societal contempt, while others we may be innately predisposed to acquire.

Throughout this chapter I have sought to embed the theory of evaluative foci in a broader multidisciplinary context, ranging from neuroscience, through artificial intelligence and animal cognition, to evolutionary theory. I aimed to show that the psychological-behavioral account of moral judgment I developed in Chapters 2 through 5 holds its ground in light of evidence from a broader range of disciplines. The purpose here was twofold: First, as I argued in Chapter 1, naturalistic theories in philosophy ought to hinge on evidence from the full range of scientific disciplines that are relevant to the question of philosophical interest. Second, I have defended a dual-process theory of moral judgment and – as I laid out in Chapter 2 – dual-process accounts must fulfill certain desiderata. In this chapter, I focused on three such features introduced in Chapter 2: parallel processing, computational load, and evolutionary primacy.

We observed that the neuroscience of moral judgment and political orientation demonstrates the double dissociation of two neurocognitive systems, corresponding to action and outcome-based moral judgments. Therefore, the individual differences in moral and political attitudes discussed in Chapter 5 may be the result of differences in the reliance on, and interplay between, these neurocognitive processes. Further, the studies on computational models of decision-making demonstrate that action-based (model-free) evaluation generally imposes a lighter computational load, and yet is more prone to yielding errors, than outcome-based evaluation. Lastly, the evolutionary considerations presented towards the end suggest that action-based evaluations of self and others in the community might have posed a few adaptive advantages on members of ancestral communities, in maintaining homogeneity and thereby managing disease threat and building social trust. Together, the computational and evolutionary perspectives strongly suggest that model-free evaluations arose as useful heuristics for the regulation of own behavior and the spontaneous evaluation of others.

In turn, this exercise revealed why reason would result in the preference for outcome- over action-based moral assessments. As most heuristics, model-free moral evaluation is prone to prediction errors when deployed in novel circumstances. In other words, in circumstances that are unusual – by comparison, either to the world in which our innate predispositions were forged or to the conditions in which we acquired action-based value associations – these value representations are likely to produce suboptimal results. Reflective individuals may be more likely to notice this tendency of our moral minds, and opt for a model-based assessment of outcomes in order to make better, more moral, decisions.

References

Adams, R. M. (2002). *Finite and infinite goods: A framework for ethics*. New York: Oxford University Press.

Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 21-62.

Alesina, A., & La Ferrara, E. (2002). Who trusts others? *Journal of Public Economics, 85*, 207-234.

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., Jost, J. T., Master, S. L., & Yee, C. M. (2007). Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience*, *10*(10), 1246-1247.

Apicella, C. L., Marlowe, F. W., Fowler, J. H., & Christakis, N. A. (2012). Social networks and cooperation in hunter-gatherers. *Nature: Letters*, *481*, 497-501.

Aristotle. (1999). *Nicomachean ethics.* T. Irwin, (ed.), London: Hackett.

Avenanti, A., Paluello, L. M., Bufalari, I., & Aglioti, S. M. (2006). Stimulus-driven modulation of motor-evoked potentials during observation of others' pain. *NeuroImage*, *32*(1), 316-324.

Ayer, A. J. (1952). *Language, truth and logic*. New York: Dover Publications.

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology, 16*(18), 1818-1823.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269-278.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From statistical formats to cognitive structures. *Behavioral and Brain Sciences,* 30(3), 287-292.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462-479.

Bargh, J. A., & Ferguson, M. L. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin*, *126*(6), 925-945.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences* 17(1): 1-10.

Baron, J. (2011). Utilitarian emotions: suggestions from introspection. *Emotion Review,* 3(3), 286-287.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*(1), 1-16.

Baron, R. M., & Kenny, D. A. (1986). Moderator-mediator variables distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-82.

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

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

Batson, C. D. (1994). Prosocial motivation: Why do we help others? In A. Tesser (Ed.), *Advanced social psychology* (pp. 333-381). Boston: McGraw-Hill.

Batson, C. D., Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E. L., Gale, S., Hassan, O. & Sampat, B. (1993). "…As you would have them do onto you": Does imagining yourself in the other's place stimulate moral action? *Personality and Social Psychology Bulletin, 29*(9), 1190-1201.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252-1265.

Baumeister, R. F., & Tierney, J. (2011). *Willpower: Rediscovering the greatest human strength*. New York: Penguin Press.

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

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293-1295.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*(2), 187-202.

Black, J. B., Turner, T. J., & Bower, G. H. (1979). Point of view in narrative comprehension, memory, and production. *Journal of Verbal Learning & Verbal Behaviour, 18,* 187–198.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition*, *57*, 1-29.

Blair, R. J. R. (2004). The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain and Cognition*, *55*, 198–208.

Blair, R. J. R. (2008). The amygdala and ventromedial prefrontal cortex: functional contributions and dysfunction in psychopathy. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, *363*(1503), 2557-2565.

Blair, R. J. R., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: a lack of responsiveness to distress cues? *Psychophysiology*, *34*, 192–198.

Botvinick, M., Jha, A. P., Bylsma, L. M., Fabian, S. A., Solomon, P. E., & Prkachin, K. M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *NeuroImage*, *25*, 312-319.

Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3-5.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306-307.

Carr, L., Iacoboni, M., Dubeau, M., Mazziotta, J., & Lenzi, G. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences*, *100*(9), 5497-5502.

Chaiken, S., & Trope, Y. (1999). *Dual process theories in social psychology*. New York: Guildford Press.

Cheng, Y., Yang, C.Y., Lin, C.P., Lee, P.R., & Decety, J. (2008). The perception of pain in others suppresses somatosensory oscillations: a magneto-encephalography study. *NeuroImage*, *40*, 1833-1840.

Chirumbolo, A. (2002). The relationship between need for cognitive closure and political orientation: The mediating role of authoritarianism. *Personality and Individual Differences*, *32*, 603-610.

Ciaramelli, E., Muccioli, M., Ladavas, E, & Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*, 84-92.

Cochin, S., Barthélémy, C., Roux, S., Martineau, J. (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience*, *11*, 1839–1842.

Copp, D. (1997). Belief, reason, and motivation: Michael Smith's ‘The moral problem’. *Ethics, 108*, 33-54.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363-366.

Crockett, M. J., & Clark, L., & Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107*(40), 17433-17438.

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

Cushman, F. A., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals impact doing/allowing judgments. *Cognition*, *108*(1), 281-289.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *7*(12), 1082-1089.

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

Dahl, A., Campos, J. J., Anderson, D. I., Uchiyama, I., Witherington, D. C., Ueno, M., Poutrain-Lejeune, L., & Barbu-Roth, M. (2013). The epigenesis of wariness of heights. *Psychological Science, 24*(7), 1361-1367.

Damasio, A. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *351*(1346), 1413-1420.

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt.

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

Daw, N., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593-620.

Dawkins, R. (1976). *The selfish gene.* New York: Oxford University Press.

Dayan, P., & Niv, Y. (2008). Reinforcement learning and the brain: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185-196.

de Waal, F. (1996). *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

de Waal, F. (2009). *Primates and philosophers: how morality evolved.* Princeton, NJ: Princeton University Press.

Deeley, Q., Daly, E., Surguladze, S., Tunstall, N., Mezey, G., Beer, D., Ambikapathy, A., Robertson, D., Giampietro, V., Brammer, M. J., Clarke, A., Dowsett, J., Fahy, T., Phillips, M. L., & Murphy, D. G. (2006). Facial emotion processing in criminal psychopathy: Preliminary functional magnetic resonance imaging study. *British Journal of Psychiatry, 189,* 533-539.

Delhey, J., & Newton, K. (2005). Predicting cross-national levels of social trust: Global pattern or nordic exceptionalism? *European Sociological Review*, *21*(4), 311-327.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*, 176-180.

Dickinson, A., Balleine, B. W., Watt, A., Gonzales, F., & Boakes, R. A. (1995). Overtraining and the motivational control of instrumental action. *Animal Learning & Behavior*, *22*, 197-206.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735-755.

Ditto, P. H., & Liu, B. (2012). Deontological dissonance and the consequentialist crutch. In M. Mikulincer, P. R. Shaver (Eds.). *The social psychology of morality: Exploring the causes of good and evil*. (pp. 51-70). Washington, D.C.: American Psychological Association.

Dungan, J., Chakroff, A., & Young, L. (in prep.). Purity versus pain: Distinct moral concerns for self versus other.

Eidelman, S., Crandall, C. S., Goodman, J. A., & Blanchar, J. C. (2012). Low-effort thought promotes political conservatism. *Personality and Social Psychology Bulletin*, *38*(6), 808-820.

Eisenberg-Berg, N. (1979). The development of children's prosocial moral judgment. *Developmental Psychology*, *15*, 128-137.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*, 709–724.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390-405.

Evans, J. S. B. T. (1989). *Biases in human reasoning: Causes and consequences*. London: Erlbaum.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove: Psychology Press.

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

Feltz, A., & Cokely, E. T. (2008). The fragmented folk: More evidence of stable individual differences in moral judgments and folk intuitions. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1771-1776). Austin, TX: Cognitive Science Society.

Fincher, C. L., & Thornhill, R. (2012). Parasite stress promotes in-group assortative sociality: the cases of strong family ties and heightened religiosity. *Behavioral and Brain Sciences*, *35*(2), 61-79.

Fischer, J. M., & Ravizza, M. (1992). *Ethics: Problems and principles*. New York: Holt, Rinehart & Winston.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77-83.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5-15.

Foot, P. (1977). Euthanasia. *Philosophy & Public Affairs*, *6*(2), 85-112.

Fotopoulou, A., Conway, M. A., & Solms, M. (2007). Confabulation: Motivated reality monitoring. *Neuropsychologia*, *45*(10), 2180-2190.

Frank, R. (1988). *Passions within reason: The strategic role of the emotions*. W.W. Norton & Company, New York.

Frankham, R. (1998). Inbreeding and extinction: Island populations. *Conservation Biology*, 12(3), 665-675.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42.

Frenkel-Brunswik, E. (1948). Tolerance toward ambiguity as a personality variable. *American Psychologist*, *3*, 268.

Freud, S. (1957). Instincts and their vicissitudes. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 14, pp. 109–140). London: Hogarth Press.

Freud, S. (1990). *New introductory lectures on psycho-analysis.* P. Gay (Ed.), New York: W.W. Norton.

Funder, D. C. (2004). *The personality puzzle* (3rd ed.). New York: W. W. Norton.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493–501.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8*(9), 396-403.

Gernsbacher, M. A., Goldsmith, H. H., & Robertson, R. R. W. (1992). Do readers mentally represent characters' emotional states? *Cognition and Emotion*, *6,* 89–111.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2012). Retrospective reevaluation in sequential decision-making: A tale of two systems. *Journal of Experimental Psychology: General,*

Gibbard, A. (1990). *Wise choices, apt feelings: A theory of normative judgment*. Cambridge, MA: Harvard University Press.

Gilovich, T. (2012, April). Intuition and reason in judgment and choice. *Social Cognitive Science Brown Bag Series*. Brown University, Providence, RI.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research*, *21*, 241-253.

Gläscher, J., Daw, N., Dayan, P., O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585-595.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.

Gopnik, A., & Schwitzgebel, E. (1998). Whose concepts are they, anyway? The role of philosophical intuition in empirical psychology. In M., DePaul and W., Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry*. Lanham, MD: Rowman & Littlefield.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, *1*, 158–171.

Gordon, R. (2004). Intentional agents like myself. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science,* Vol. 2, Cambridge, MA: MIT Press.

Grafman, J., Schwab, K., Warden, D., Pridgen, A., Brown, H. R., & Salazar, A. M. (1996). Frontal lobe injuries, violence, and aggression: a report of the Vietnam head injury study. *Neurology*, *46*(5), 1231-1238.

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives use different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029-1046.

Graham, J., Sherman, G., Iyer, R., Hawkins, C.B., Haidt, J., & Nosek, B.A. (in preparation). Political ideology moderates nonpolitical moral decision-making processes, in preparation.

Gray, K., Young, L., Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101-124.

Gray, K., & Wegner, D. M. (2008). The sting of intentional pain. *Psychological Science, 19,* 1260-1262.

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 3: The neuroscience of morality*. *Emotion, disease, and development*. Cambridge, MA: MIT Press.

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

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science*, *6*, 517-523.

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., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences USA*, *106*(30), 12506-12511.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral Judgment. *Science*, *293*, 2105-2108.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814:34.

Haidt, J. (2012). *The righteous mind: Why good peoploe are divided by politics and religion*. New York: Pantheon.

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 2: The cognitive science of morality: Intuition and diversity*. Cambridge, MA: MIT Press. (pp. 181-217).

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

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize, *Social Justice Research*, *20*, 98-116.

Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In J. Jost, A. C. Kay & H. Thorisdottir (Eds.), *Social and psychological bases of ideology and system justification* (pp. 371-401). New York: Oxford University Press.

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

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.

Haidt, J., McCauley, C., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences*, *16*, 701-713.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: infants prefer those who harm dissimilar others. *Psychological Science*, *24*(4), 589-594.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557-559.

Hare, R. M. (1952). *The language of morals*. Oxford: Clarendon Press.

Harrison, N. A., Singer, T., Rotshtein, P., Dolan, R. J., & Critchley, H. D. (2006). Pupillary contagion: Central mechanisms engaged in sadness processing. *Social, Cognitive and Affective Neuroscience*, *1*, 5-17.

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

Heaven, P. C. L., & Ciarrochi, J., & Leeson, P. (2011.) Cognitive ability, right-wing authoritarianism, and social dominance orientation: A five-year longitudinal study amongst adolescents. *Intelligence*, *39*(1), 15-21.

Helweg-Larsen, M., & Nielsen, G. A. (2009). Smoking cross-culturally: Risk perceptions among young adults in Denmark and the United States. *Psychology and Health*, *24*(1), 81-93.

Helzer, E. G., & Pizarro, D. A. (2011). Dirty liberals! Reminders of physical cleanliness influence moral and political attitudes. *Psychological Science*, *22*(4), 517-522.

Henrich, J. (2004). Cultural group selection, coevolutionary processes, and large-scale cooperation. *Journal of Economic Behavior & Organization, 53,* 3-143.

Hodson, G., &. Busseri, M. A. (2012). Bright minds and dark attitudes: Lower cognitive ability predicts greater prejudice through right-wing ideology and low intergroup contact. *Psychological Science*, *23*(2), 187-195.

Hoffman, M. L. (1982). Development of prosocial motivation: empathy and guilt. In N. Eisenberg (Ed.), *Development of Prosocial Behavior* (pp. 281–313). New York: Academic Press.

Hoffman, M. L. (2000). *Empathy and moral development: implications for caring and justice*. Cambridge, UK: Cambridge University Press.

Holmes, D. S. (1978). Projection as a defense mechanism. *Psychological Bulletin, 85*(4), 677-688.

Horberg, E.J., Oveis, C., & Keltner, C. (2011). Emotions as moral amplifiers: An appraisal tendency approach to influences of distinct emotions upon moral judgment. *Emotion Review 3*, 237-244.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology, 97,* 963-976.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14,* 399-425.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, 1*3(*1), 1-6.

Hume, D. (1739). *A treatise of human nature*: *Being an attempt to introduce the experimental method of reasoning into moral subjects*. London: John Noon.

Hume, D. (1751/1894). *An enquiry concerning the principles of morals*. Oxford: Clarendon Press.

Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*, 2526–8.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009a). Conservatives are more easily disgusted than liberals. *Cognition and Emotion*, *23*, 714-725.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012). Disgusting smells cause decreased liking of gay men. *Emotion*, *12*, 23-27.

Inbar, Y., Pizarro, D. A., Knobe, J., & Bloom, P. (2009b). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, *9*(3), 435-439.

Jackson, P. L., Brunet, E., Meltzoff, A. N., & Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain. *Neuropsychologia*, *44*, 752-761.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage*, *24*, 771–9.

James, W. (1890/1950). *Principles of psychology*. New York: Dover.

Janoff-Bulman, R., Sheikh, S., & Baldacci, K. (2008). Mapping moral motives: Approach, avoidance, and political orientation. *Journal of Experimental Social Psychology*, *44*(4), 1091-1099.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sullaway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339-375.

Jost, J. T., Kruglanski, A. W., & Simon, L. (1999). Effects of epistemic motivation on conservatism, intolerance, and other system justifying attitudes. In L. Thompson, D. M. Messick, & J. M. Levine (Eds.), *Shared cognition in organizations: The management of knowledge* (pp. 91-116). Mahwah, NJ: Erlbaum.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T., Gilovich, D., Griffin, D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, UK: Cambridge University Press.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K.J. Holyoak & R.G. Morrison (eds.), *The Cambridge handbook of thinking and reasoning*. Cambridge University Press. 267-293.

Kanai, R., Feilden, T., Firth, C., & Rees, G. (2011). Political orientations are correlated with brain structure in young adults. *Current Biology*, *21*(8), 677-680.

Kant, I. (1785/1964). *Groundwork of the metaphysic of morals*. H. J. Paton (trans.), New York: Harper & Row.

Kass, L. R. (1989). Neither for love nor money: Why doctors must not kill. *Public Interest*, *94*, 25-46.

Katz, L. D. (2000). *Evolutionary origins of morality: Cross-disciplinary perspectives*. Exeter, UK: Imprint Academic.

Kelly, D. (2013). *Yuck! The nature and moral significance of disgust*. Cambridge, MA: MIT Press.

Kiehl, K. A., Smith, A. M., Hare, R. D., Mendrek, A., Forster, B. B., Brink, J., & Liddle, P. F. (2001). Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. *Biological Psychiatry*, *50*, 677–684.

Kim, J. (1993). *Supervenience and mind: selected essays*. Cambridge: Cambridge University Press.

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708-714.

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

Koeslag, J. H. (1990). Koinophilia groups sexual creatures into species, promotes stasis, and stabilizes social behaviour. *Journal of Theoretical Biology*, *144*, 15-35.

Koeslag, J.H. (1997). Sex, the prisoner's dilemma game, and the evolutionary inevitability of cooperation. *Journal of Theoretical Biology*, *189*, 53-61.

Kornblith, H. (1998). The role of intuition in philosophical inquiry: An account with no unnatural ingredients. In M., DePaul and W., Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry*. Lanham, MD: Rowman & Littlefield.

Korsgaard, C. (1996). *The sources of normativity*. Cambridge, UK: Cambridge University Press.

Kruglanski, A. W. (2005). *The psychology of closed-mindedness*. New York: Psychology Press.

Kruglanski, A. W., & Orehek, E. (2007). Partitioning the domain of social inference: Dual mode and systems models and their alternatives. *Annual Review of Psychology*, 58, 291-316.

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology*, *65*(5), 861-876.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month olds. *Psychological Science*, *14*(5), 402-408.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*, *19*, 42-58.

Lamm, C., Nusbaum, H.C., Meltzoff, A.N., & Decety, J. (2007). What are you feeling? Using functional magnetic resonance imaging to assess the modulation of sensory and affective responses during empathy for pain. *PLoS ONE*, *12*, e1292.

Langlois, J. H., & Roggman, L. (1990). Attractive faces are only average. *Psychological Science*, *1*, 115-121.

László, J., & Cupchik, G. C. (1995). The role of affective processes in reading time and time experience during literary reception. *Empirical Studies of the Arts*, *13*, 25–37.

LeDoux, J. (1998). Fear and the brain: where have we been, and where are we going? *Biological Psychiatry*, *44*(12), 1229-1238.

Li, X., Lu, Z-L., D'Argembeau, A., Ng, M., & Bechara, A. (2010). The Iowa gambling task in fMRI images. *Human Brain Mapping*, *31*, 410–423.

Lieberman, D., & Lobel, T. (2012). Kinship on the Kibbutz: coresidence duration predicts altruism, personal sexual aversions and moral attitudes among communally reared peers. *Evolution and Human Behavior*, *33*(1), 26-34.

Lieberman, D., Tooby, J. & Cosmides, L. (2007). The architecture of human kin detection. *Nature*, *445*, 727-731.

LoBue, V., & DeLoache, J. S. (2008). Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science*, *19*(3), 284-289.

Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. New York: Pelican.

Madestam, A., & Yanagizawa-Drott, D. (2012). Shaping the nation: The effect of Fourth of July on political preferences and behavior in the United States. *HKS Faculty Research Working Paper Series* RWP12-034.

Mar, R. (2004). The neuropsychology of narrative: story comprehension, story production and their interrelation. *Neuropsychologia*, *42*, 1414–1434.

Matthews, M., Levin, S., & Sidanius, J. (2009). A longitudinal test of the model of political conservatism as motivated social cognition. *Political Psychology, 30,* 921-936.

Mikhail, J. (2000). Rawls' linguistic analogy: A study of the ''generative grammar" model of moral theory described by John Rawls in a theory of justice, unpublished doctoral dissertation. Cornell University.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and future. *Trends in Cognitive Sciences*, *11*(4), 143-152.

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 3: The neuroscience of morality*. Cambridge, MA: MIT Press.

Mikhail, J. (2011). Emotion, neuroscience, and law: A comment on Darwin and Greene. *Emotion Review -- Special Issue: Emotion and Morality*, *3*(3), 293-295.

Mill, J.S. (1863). *Utilitarianism*. London: Parker, Son, & Bourn.

Miller, R. M., & Cushman, F. A. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass*, *7*(10), 707–718.

Miller, R. M., Hannikainen, I., & Cushman, F. A. (2013). Bad actions or bad outcomes? Differentiating affective contributions to the condemnation of harm. *Emotion*, in press.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, *50*, 655-663.

Moll, J., Eslinger, P. J., & de Oliveira-Souza, R. (2001). Frontopolar and anterior temporal cortex activation in a moral judgment task: preliminary functional MRI results in normal subjects. *Arquivos de Neuro-Psiquiatria*, *59*(3B), 657-664.

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002a) Functional networks in emotional moral and nonmoral social judgments. *Neuroimage*, *16* (3), 696-703.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002b). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience*, *22*(7), 2730-2736.

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,* 549-557.

Moriguchi, Y., Decety, J., Ohnishi, T., Maeda, M., Mori, T., Nemoto, K., Matsuda, H., & Komaki, G. (2007). Empathy and judging others' pain: an fMRI study of alexithymia. *Cerebral Cortex*, *17*(9), 2223-2234.

Morrison, I., Lloyd, D., di Pellegrino, G., & Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex: Is empathy a multisensory issue? *Cognitive, Affective & Behavioral Neuroscience*, *4*(2), 270-278.

Murray, D. R., Schaller, M., & Suedfeld, P. (2013). Pathogens and politics: Further evidence that parasite prevalence predicts authoritarianism. *PLoS ONE*, *8*(5), e62275.

Narayan, V. M., Narr, K. L., Kumari, V., Woods, R. P., Thompson, P. M., Toga, A. W., & Sharma, T. (2007). Regional cortical thinning in subjects with violent antisocial personality disorder or schizophrenia. *American Journal of Psychiatry*, *164*(9), 1418-27.

Nichols, S. (2002). Norms with feeling: towards a psychological account of moral judgment. *Cognition*, *84*, 221-236.

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

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*, 530-542.

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

Oatley, K. (1999). Why fiction may be twice as true as fact: Fiction as cognitive and emotional simulation. *Review of General Psychology*, *3,* 101–117.

Ogino, Y., Nemoto, H., Inui, K., Saito, S., Kakigi, R., & Goto, F. (2007). Inner experience of pain: imagination of pain while viewing images showing painful events forms subjective pain representation in human brain. *Cerebral Cortex*, *17*, 1139-1146.

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483-522.

Olatunji, B. O., Williams, N. L., Tolin, D. F., Sawchuk, C. N., Abramowitz, J. S., Lohr, J. M., & Elwood, L. (2007). The Disgust Scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment*, *19*, 281-297.

Olson, L. R., & Green, J. C. (2006). The religion gap. *Political Science & Politics*, *39*(3), 455-459.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement learning systems by taxing the central executive. *Psychological Science*. Advance online publication.

Özyürek, A., & Trabasso, T. (1997). Evaluation during the understanding of narratives. *Discourse Processes*, *23*, 305–355.

Papineau, D. (1993). *Philosophical naturalism*. Oxford: Blackwell.

Paxton, J. M., Ungar, L., and Greene, J.D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163-177.

Perry, R., & Sibley, C. G. (2012). Big-Five personality prospectively predicts social dominance orientation and right-wing authoritarianism. *Personality and Individual Differences 52*(1), 3-8.

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.

Pew Research. (2008, Dec 2). Americans say they like diverse communities; election, census trends suggest otherwise. *Pew Research: Social and Demographic Trends*. Retrieved December 29 2013 from http://www.pewsocialtrends.org/2008/12/02/americans-say-they-like-diverse-communities-election-census-trends-suggest-otherwise/.

Piazza, J. (2012). If you love me keep my commandments: Religiosity increases preference for rule-based moral arguments. *International Journal for the Psychology of Religion*, *22*(4), 285-302.

Piazza, J., & Sousa, P. (2013). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*.

Pinillos, N., Smith, N., Nair, G., Marchetto, P., & Mun, C. (2011). Philosophy's new challenge: Experiments and intentional action. *Mind & Language*, *26*(1), 115-139.

Pizarro, D. A. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behavior*, *30*(4), 355-375.

Plato. (1991). *The republic*. Allan Bloom (ed.), New York: Basic Books.

Preston, S. D., & de Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, *25*(1), 1-20.

Prinz, J. (2006). The emotional basis of moral judgment. *Philosophical Explorations, 9*, 29–43.

Prinz, J. (2007). Is morality innate? In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 1: The evolution of morality. Adaptations and innateness*. Cambridge, MA: MIT Press.

ProCon.Org. (2011, Aug 17). Euthanasia: opinion polls. Retrieved January 31 2014 from (http://euthanasia.procon.org/view.resource.php?resourceID=000134).

Putnam, R. (2007). E pluribus unum: Diversity and community in the twenty-first century: The 2006 Johan Skytte Prize lecture. *Scandinavian Political Studies, 30*(2), 137-174.

Rachels, J. (1986). *The end of life: Euthanasia and morality*. Oxford: Oxford University Press.

Raine, A., Lencz, T., Bihrle, S., LaCasse, L., & Colletti, P. (2000). Reduced prefrontal graymatter volume and reduced autonomic activity in antisocial personality disorder. *Archives of General Psychiatry*, *57*, 119–129.

Rall, J., & Harris, P. L. (2000). In Cinderella’s slippers? Story comprehension from the protagonist’s point of view. *Developmental Psychology*, *26*, 202–208.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. (2009). Positive interactions promote public cooperation. *Science*, *325*(5945), 1272-1275.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*, 427-430.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.

Ritov, I., & Baron, J. (1995). Outcome knowledge, regret, and omission bias. *Organizational Behavior and Human Decision Processes*, *64*, 119-127.

Ritov, I., & Baron, J. (1999). Protected values and omission bias: *Organizational Behavior and Human Decision Processes*, *79*, 79-94.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169-192.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996). Localization of grasp representation in humans by PET: 1. Observation versus execution. *Experimental Brain Research*, *111*(2), 246–252.

Ross, L., Greene, D., & House, P. (1977). The ‘false consensus effect’: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279-301.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165-184.

Rozin, P., & Singh, L. (1999). The moralization of cigarette smoking in the United States. *Journal of Consumer Psychology*, *8*(3), 321-337.

Rozin, P., Haidt, J., McCauley, C., Dunlop, L., & Ashmore, M. (1999a). Individual differences in disgust sensitivity: Comparisons and evaluations of paper-and-pencil versus behavioral measures. *Journal of Research in Personality*, *33*, 330-351.

Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotions*, 3rd ed. (pp. 757-776). New York: Guilford Press.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999b). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral ethics (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574-586.

Rozin, P., Markwith, M., & Stoess, C. (1997). Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological Science*, *8*(2), 67-73.

Rubenstein, A. J., Kalakanis, L., & Langlois, J. H. (1999). Infant preferences for attractive faces: A cognitive explanation. *Developmental Psychology*, *35*(3), 848-855.

Saad, L. (2012, August 22). One in five U.S. adults smoke, tied for all-time low. *Gallup: Well Being*. Retrieved December 29, 2013, from http://www.gallup.com/poll/156833/one-five-adults-smoke-tied-time-low.aspx.

Saarela, M. V., Hlushchuk, Y., Williams, A. C., Schurmann, M., Kalso, E., & Hari, R. (2007). The compassionate brain: Humans detect intensity of pain from another's face. *Cerebral Cortex*, *17*(1), 230-237.

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*, 1529-1546.

Schnall, S., Haidt, J., Clore, G. L., Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality & Social Psychology Bulletin*, *34*(8), 1096-1109.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1599.

Searle, J. R. (1964). How to derive 'ought' from 'is'. *Philosophical Review*, *73*(1), 43–58.

Sellars, W. (1956). Empiricism and the philosophy of mind. In Herbert Feigl and Michael Scriven (Eds.), *Minnesota studies in the philosophy of science, vol. 1: The foundations of science and the concepts of psychology and psychoanalysis*, (pp. 253-329).

Shafir, E., & LeBouef, R. A. (2002). Rationality. *Annual Review of Psychology, 53*, 491-517.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667-677.

Shenhav, A., Rand, D., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423-428.

Shepher, J. (1983). *Incest: a biosocial view*. New York: Academic Press.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The “big three” of morality (autonomy, community, and divinity), and the "big three" explanations of suffering. In Allan Brandt and Paul Rozin (Eds.), *Morality and health*. New York: Routledge.

Singer, P. (1981). *The expanding circle: ethics and sociobiology*. New York: Farrar, Straus and Giroux.

Singer, P. (1995). *Rethinking life & death: The collapse of our traditional ethics*. Oxford: Oxford University Press.

Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157-1162.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3-22.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D.G. (2002).The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman, (Eds.), *Intuitive judgment: Heuristics and biases. Cambridge University Press*.

Smetana, J. G. (1985). Preschool children’s conceptions of transgressions: Effects of varying moral and conventional domain-related attributes. *Developmental Psychology*, *21*(1), 18-29.

Smith, E. R., & DeCoster, J. (2000). Dual process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108-131.

Smith, A. (1759/2011). *The theory of moral sentiments.* New York: Gutenberg Publishers.

Smith, M. (1987). The Humean theory of motivation. *Mind*, 96(381), 36-61.

Sober, E., & Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General. 127*, 161-188.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645-726.

Stanovich, K. E. (2011). *Rationality and the reflective mind.* New York: Oxford University Press.

Stevenson, C. L. (1944). *Ethics and language.* New Haven, CT: Yale University Press.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill-learning: A dual-process approach. *Psychological Review*, *112*(1), 159-192.

Sunstein, C. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531-542.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*, 9-44.

Sutton, R. S., & Barto, A. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, *11*(1), 126-134.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*(6), 1395-1415.

Tiihonen, J., Hodgins, S., Vaurio, O., Laakso, M., Repo, E., Soininen, H., Aronen, H. J., Nieminen, P., & Savolainen, L. (2000). Amygdaloid volumen loss in psychopathy. *Society for Neuroscience Annual Meeting*. New Orleans, LA (USA).

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*, 455-460.

Trivers, R. L. (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*. Cambridge, MA: Cambridge University Press.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476-477.

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: a meta-analysis. *Cognitive, Affective and Behavioral Neuroscience, 3*(4), 255-274.

Wegner, D. M., & Bargh. J. A. (1998). Control and automaticity in social life. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 446-496.) New York: McGraw-Hill.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67,* 1049-1062.

Wegner, D. M., & Giuliano, T. (1980). Arousal-induced attention to self. *Journal of Personality and Social Psychology, 38*(5), 719-726.

Wegner, D. M., & Giuliano, T. (1983). Social awareness in story comprehension. *Social Cognition, 2,* 1-17.

Westermarck, E. (1891). *The history of human marriage*. London: Macmillan and Co.

Wheatley, T., & Haidt, J. (2005). Hypnotically induced disgust makes moral judgments more severe. *Psychological Science, 16,* 780-784.

Wilson, D. S., & Wilson, E. O. (2008). Evolution 'for the good of the group'. *American Scientist, 96,* 380-389.

Wilson, G. D. (1973). *The psychology of conservatism*. London: Academic Press.

Wolf, A. P., & Durham, W. H. (2004). *Inbreeding, incest, and the incest taboo: The state of knowledge at the turn of the century*. Stanford, CA: Stanford University Press.

Wolf, A. P. (1995). *Sexual attraction and childhood association: A Chinese brief for Edward Westermarck*. Stanford, CA: Stanford University Press.

Yang, Y., Raine, A., Narr, K. L., Colletti, P., & Toga, A. W. (2009). Localization of deformations within the amygdala in individuals with psychopathy. *Archives of General Psychiatry, 66*(9), 986-994.

Zajonc, R. (1980). Feeling and thinking: preferences need no inferences. *American Psychologist, 35*, 151-175.

Appendices

Appendix A

Feces

Gary carefully smears his own feces all over his body. He looks at himself in the mirror and thoroughly washes himself off after a short while.

Comical portrait

An artist paints a comical portrait of a religious icon completely naked. She never shows anyone the portrait and throws it away.

Kiss uncle

A young man French-kisses his uncle at a family party. Everyone at the party, including the young man and his uncle, finds it very funny.

Urine

Brooke stands on the roof of her house and urinates on her best friend, who is standing below, at her friend’s request.

Pubic hair sandwich

Ken shaves his pubic hair, gathers it and eats it in a sandwich. He enjoys the taste but decides never to do it again.

Two-inch tail

Carly gets plastic surgery that adds a two-inch tail to the end of her spine.

Sell soul

An old woman signs a piece of paper that says “I hereby sell my soul, after my death, to whoever has this piece of paper”. She makes a paper plane out of the note and flicks it out her apartment window.

Chicken

A man goes to the supermarket once a week and buys a whole chicken. Before cooking it, he has sexual intercourse with the chicken carcass. Then he cooks it and eats it.

Appendix B

Submarine (Personal)

Gary is the captain of a military submarine. An onboard explosion has caused Gary to lose most of the air supply and has injured one of Gary's crew who is quickly losing blood. The injured crew member is probably going to die from his wounds no matter what happens. There isn't enough air for the whole crew. The only way to save the other crew members is to shoot dead the injured crew member so that there will be just enough air for Gary and the rest of the crew to survive.

Gary shoots the injured crew member. The injured crew member dies and there is enough air for Gary and the rest of the crew to survive.

Virus (Personal)

A deadly virus is spreading around the world, killing thousands of people. Kevin is a scientist. He has invented two chemicals. One of them kills the virus. The other is a deadly poison. Kevin doesn't know which chemical is which because the labels on his containers got mixed up. Once he figures out which is which Kevin can use the good chemical to save thousands of lives, but the only way to find out is to test each chemical on someone. If Kevin does this, one of the people will die, but he will then be able to start saving many other people's lives.

Kevin tests both chemicals on his assistants. One of his assistants dies within minutes. Kevin figures out which chemical is the poison and which is the good chemical. He delivers the chemical in time to save many other people's lives.

Jungle (Personal)

Anne is part of a group studying animals in the jungle. The group includes eight children. Anne's group is captured by terrorists who keep you all locked up for several days. One of the terrorists likes Anne. He tells Anne that the leader plans to kill everyone the next day. He says he will help Anne escape, under one condition. To make sure she won't go to the police, the terrorist wants to videotape Anne killing one of the other adults. If Anne accepts his offer, the other adults will die but Anne and the eight children will escape.

Anne accepts the offer. She kills one of the other adults while the terrorists videotape her doing it. The terrorists let her and the eight children go.

Lifeboat (Personal)

Brooke is on a big boat at sea. There is a fire on the boat and everyone has to get off. People get into the lifeboats. All the lifeboats, including Brooke's, have too many people in them. The sea is getting rough, and water is coming in over the sides. If nothing is done, the lifeboat will sink and everyone on board will die. However, there is an injured man who will not survive in any case. If Brooke throw that man overboard the boat will stay afloat and the remaining passengers will be saved.

Brooke pushes the injured man off the lifeboat and he drowns in the sea. The lifeboat is now light enough to float and the remaining passengers are saved.

Climbers (Personal)

Ken is the leader of a mountain climbing group that is stranded in the wilderness. Ken's group includes a family of six that needs a certain vitamin very badly. Some people's kidneys contain large amounts of this vitamin. There is one person in Ken's party who has the vitamins in his kidneys. The only way to save the lives of the six members of this family is to remove one of this man's kidneys and give it to the six people. The man will not die, but his health will suffer. He is opposed to this plan, but Ken has the power to do as he sees fit.

Despite the man's refusal, Ken removes one of the man's kidneys. The man suffers from the operation but, thanks to the man's kidney, the family members survive.

Enemy Doctor (Personal)

The enemy has taken over Carly's village. Carly has two children, ages five and eight. There is an enemy doctor who performs painful experiments on humans that always lead to death. He intends to perform experiments on one of Carly's children, but he will allow Carly to choose which of his children he will experiment on. If Carly refuses to bring one of his children to him, he will find them both and experiment on both of them.

Carly agrees to bring one of her children to the doctor in order to save her other child. The doctor experiments on Carly's child and he dies. The other child is saved.

Vaccine (Impersonal)

Gary works for the government's public health office. Scientists have made a new vaccine to fight a deadly disease, and Gary must decide whether the government will tell people to use it. The vaccine works well to prevent the deadly disease, and it will save many lives if the vaccine is distributed. However, a small number of people will be killed by an allergic reaction with the vaccine itself. There is no way to predict who will have this reaction. So, if Gary agrees to distribute the vaccine, some people will surely be killed by it.

Gary agrees to distribute the vaccine and the vaccine saves many lives of people fighting the deadly disease. The vaccine also kills a few people who are allergic to it.

Bombing campaign (Impersonal)

Kevin has been appointed general of an army during a large war that has lasted several years. Kevin's army has gradually gained ground, and he is finally about to win. Now, Kevin has to decide his closing strategy. He could order large-scale bombing of the opposing side's home country, which will defeat their army quickly, but will also lead to many unavoidable civilian deaths. Or, he could order a ground war in which these civilian deaths will be avoided. However, the enemy will force more of its own civilians into combat during the long ground war, which will ultimately lead to more deaths than the quicker bombing campaign.

Kevin orders the quicker bombing campaign. As expected, the enemy admits defeat and the war is brought to an end but many civilians die as a result of the bombing.

Explosives (Impersonal)

Anne is the mayor of a large city. A deranged man has set up explosives in two large buildings in Anne's city and come to her office with the detonator. Both buildings have thousands of people working in them who would be killed. The deranged man has tied Anne up so that she cannot move. He intends to blow up one of the buildings, but he will allow Anne to flip a coin to determine which of the two buildings he will blow up. If Anne refuses to flip the coin, he will blow up both buildings, which will result in more total deaths.

Anne agrees to flip the coin. The deranged man blows up one of the buildings and deactivates the bomb in the other building.

Gang violence (Impersonal)

Brooke is a member of Congress. She and the other members of congress are deciding whether to adopt a new social policy intended to decrease organized crime and gang violence in urban and rural areas. The new policy would increase the number of raids on the homes of gang members and members of crime families in order to round these suspects up. There are sure to be some innocent civilians killed in these raids during gun battles with the police, but by rounding these criminals up, many more innocent civilians will be saved.

Brooke casts her vote in favor of the new social policy. The policy is passed. Organized crime and gang violence are reduced significantly but the police report some innocent civilians are killed in these raids.

Gas leak (Impersonal)

Ken is the late-night watchman in a hospital. Due to an accident in a factory next door, deadly gas is coming into the hospital. In one room of the hospital, there are three patients. In another room, there are seven patients. If Ken does nothing, the fumes will reach the three patients and kill them. The only way to save the three patients is to close a vent so that the fumes go into the room with the seven patients, but then the seven patients will die.

In order to save the seven patients, Ken leaves the vent open. The deadly gas travels to the room with three patients and the three patients die. The seven patients in the other room survive.

Motorboat (Impersonal)

Carly is driving a motorboat in the bay when she notices a couple of swimmers in trouble. The two swimmers are drowning at the end of a channel in front of her, and they need help immediately to survive. As she boats towards them, she sees another swimmer drowning much closer, not far from the side of her boat. If Carly stops to save the one swimmer to the side, she will not be able to get to the two other swimmers in time to save them. If she continues to speed towards the two swimmers, however, the one swimmer beside her will drown.

Carly decides to speeds towards the two swimmers. She reaches them in time to save them, but in the meantime the one swimmer drowns.

Appendix C

Feces (Agent)

Gary works at an adult video store. On his way to work today, Gary grabs a bagel and a cup of coffee. Once he arrives at the store, he puts his coat in the back room and then heads out front to help some customers.

Around midday there is a lull in customers and Gary decides to go and grab a sandwich for lunch from the café across the street. When he returns to the video store he has a conversation with one of his co-workers to pass the time. After the conversation, Gary decides that he will fulfill his fantasy of smearing feces all over his body after the store closes at 9 pm.

Closing time arrives and after making sure all of the customers have left, Gary’s co-worker counts the register and locks up the store. Meanwhile, Gary goes to the bathroom and defecates.

Feces (Bystander)

Steve works at an adult video store. On his way to work today, Steve grabs a bagel and a cup of coffee. Once he arrives at the store, he puts his coat in the back room and then heads out front to help some customers.

Around midday there is a lull in customers and Steve decides to go and grab a sandwich for lunch from the café across the street. When he returns to the video store he has a conversation with one of his co-workers to pass the time. After the conversation, his co-worker says that he will fulfill his fantasy of smearing feces all over his body after the store closes at 9 pm.

Closing time arrives and Steve’s co-worker goes to the bathroom and defecates. After making sure all of the customers have left, Steve counts the register and locks up the store.

Blasphemy (Agent)

Julie is an artist who lives in the heart of Boston with her husband and two children. One spring morning, Julie wakes up early to meet a friend and fellow artist at a nearby café for breakfast. Julie decides to buy a cup of coffee and a blueberry muffin before she sits down with her friend.

They discuss collaborating on a new project and a proposed budget for the project. On the way to the art studio, Julie notices city staff putting up decorations to celebrate the upcoming religious holiday. When they arrive at the art studio, Julie's friend begins to brew a pot of tea.

As her friend is making the tea, Julie prepares her workstation and says she wants to draw a portrait of the religious icon she saw on the way to the studio. However, she would depict him nude and make the portrait rather comical with the religious leader's genitalia appearing grossly enlarged.

Blasphemy (Bystander)

Meg is an artist who lives in the heart of Boston with her husband and two children. One spring morning, Meg wakes up early to meet a friend and fellow artist at a nearby café for breakfast. Meg decides to buy a cup of coffee and a blueberry muffin before she sits down with her friend.

They discuss collaborating on a new project and a proposed budget for the project. On the way to the art studio, Meg's friend notices city staff putting up decorations to celebrate the upcoming religious holiday. When they arrive at the art studio, Meg starts brewing a pot of tea.

As she is making the tea, Meg's friend prepares her workstation and talks about wanting to draw a portrait of the religious icon she saw on the way to the studio. However, she says that she would depict him nude and make the portrait rather comical with the religious leader's genitalia appearing grossly enlarged. After listening to her friend talk, Meg pours herself some tea and sits down to work.

Chicken (Agent)

Bill is a middle-aged postman and delivers mail to the residents of a small town in Michigan. On his way home, Bill goes to the supermarket to buy his weekly groceries where he runs into his friend. Bill invites his friend over for beers and a game of poker that same night. Bill hadn't seen his friend in some time and was very excited for the night to begin.

Bill's friend arrives at Bill's house and they play few games of poker and drink a few beers. Afterwards, Bill and his friend get to talking about sex and discussing some of their most exciting sexual experiences.

Bill's friend says "It's getting late; I should head home." Once his friend has left, Bill realizes that he is hungry and decides to roast the whole chicken that he bought at the supermarket. However, before cooking it, Bill would like to have sex with the chicken carcass.

Chicken (Bystander)

Evan is a middle-aged postman and delivers mail to the residents of a small town in Michigan. On his way home, Evan goes to the supermarket to buy his weekly groceries where he runs into his friend. Evan's friend invites him over for beers and a game of poker that same night. Evan hadn't seen his friend in some time and was very excited for the night to begin.

Evan arrives at his friend's house and they sit down to play few games of poker and drink a few beers. Afterwards, Evan and his friend get to talking about their personal lives and discussing some of their most exciting sexual experiences.

Evan feels that it's getting late and that he should go home. Once Evan has left, Evan's friend decides to roast the whole chicken that he bought at the supermarket. However, before cooking it, Evan's friend decides to have sex with the chicken carcass.

Incest (Agent)

Eric is a businessman and has a very hectic schedule. Usually he is on top of things, but this week was really intense for him. Eric wakes up on Saturday morning and is happy that it is the weekend. However, he soon realizes that he hasn't gotten a present to take to his cousin's birthday party that afternoon.

Eric hastily throws on some clothes and heads to a nearby Target where he buys a gift. Eric rushes to the car and drives to the party. Many of Eric's family members are at the party and once he arrives they all sit down to a big barbecue lunch. The lunch is very enjoyable, filled with talking, laughter, and plenty of margaritas.

After lunch is finished, Eric's sister puts on some music and everyone starts to dance and continue to drink. His sister then starts to play a game of cards with some of her relatives. Out of the blue, Eric exclaims that he wants to passionately French kiss his uncle.

Incest (Bystander)

Sally is a businesswoman and has a very hectic schedule. Usually she is on top of things, but this week was really intense for her. Sally wakes up on Saturday morning and is happy that it is the weekend. However, she soon realizes that she hasn't gotten a present to take to her cousin's birthday party that afternoon.

Sally hastily throws on some clothes and heads to a nearby Target where she buys a gift. Sally rushes to the car and drives to the party. Many of Sally's family members are at the party and once she arrives they all sit down to a big barbecue lunch. The lunch is very enjoyable, filled with talking, laughter, and plenty of margaritas.

After lunch is finished, Sally puts on some music and everyone starts to dance and continue to drink. Out of the blue, her brother exclaims that he wants to passionately French kiss their uncle while Sally is playing a game of cards with some of her relatives.

Sell Soul (Agent)

Martha is an old woman who lives alone in an apartment in Brooklyn, New York. One day, Martha goes online and orders a book about witchcraft. The day the book is delivered to her apartment, Martha invites her niece over for lunch.

When her niece arrives, Martha tells her about the new book she bought. She tells her niece that she wants to sell her soul and that the book explains how to do it. Martha then walks over to the dresser, grabs a piece of paper, and shows it to her niece. Martha takes back the paper and reads the note, "I hereby sell my soul, after my death, to whoever has this piece of paper". Martha begins to explain her plan as her niece finishes setting the table for lunch. Martha says she wants to sign the bottom, put it in a sealed envelope and leave it on the steps of a haunted warehouse.

Sell Soul (Bystander)

Lily is a young woman who lives alone in an apartment in Brooklyn, New York. One day, while Lily is working at her computer, she gets a call from her aunt asking if she would come over for lunch later that day.

Lily arrives at her aunt's house and her aunt tells her about the new book on witchcraft she bought. She tells Lily that she wants to sell her soul and that the book explains how to do it. Lily's aunt then walks over to the dresser, grabs a piece of paper, and shows it to Lily.

Lily reads the note, "I hereby sell my soul, after my death, to whoever has this piece of paper". She gives the piece of paper back to her aunt and her aunt says she wants to sign the bottom, put it in a sealed envelope and leave it on the steps of a haunted warehouse. As her aunt is speaking, Lily finishes setting the table for lunch.

Urine (Agent)

Brooke is a 17 year-old student at a prep school in Providence, Rhode Island. One weekend, Brooke's parents leave town to celebrate their anniversary at the hotel where they were married. Brooke's parents are usually very strict and don't allow her to have friends over. Brooke decides to capitalize on her parents' absence and invites her two best friends over for dinner and a movie.

As they make dinner together, the girls discuss some of the hot gossip going around their school. After dinner, they go to the couch and watch a movie. Brooke says that the night has been amazing so far, but suggests they head up to the roof to make things even more exciting.

Brooke's friends agree and follow her to the roof, where they sit for a long time and watch the stars. As they are talking, Brooke realizes she has to pee and tells her friends that she needs to use the bathroom. Brooke's friend Anna confesses that she has always wanted to be urinated on. Brooke thinks this is a good idea and tells Anna that she would be willing to urinate on her.

Urine (Bystander)

Ellie is a 17 year-old student at a prep school in Providence, Rhode Island. One weekend, Ellie gets a call from her best friend. Her friend's parents are out of town for their anniversary and she wanted to know if Ellie could come over to her place for dinner and a movie. She tells Ellie to bring their other friend Anna with her.

As they make dinner together, the girls discuss some of the hot gossip going around their school. After dinner, they go to the couch and watch a movie. One of Ellie's friends says that the night has been amazing so far, but suggests they head up to the roof to make things even more exciting.

Ellie agrees and follows her friends to the roof, where they sit for a long time and watch the stars. As the girls are talking, Ellie's best friend exclaims that she needs to pee. Anna then confesses that she has always wanted to be urinated on. Ellie's friend says she thinks this is a good idea and would be willing to urinate on Anna.

Hair sandwich (Agent)

After a two-hour chemistry lecture, Ken and his roommate head back to their dorm room. When they arrive at the room they throw their backpacks in a corner, jump on the couch, and turn on the TV, which goes straight to the Food Network.

Ken is hungry and sees that the special is on rhubarb and cauliflower. “Delicious!” exclaims his roommate. Ken frowns and says, “That’s disgusting! I would rather eat my own pubic hair.” As the episode continues, Ken starts to get really hungry and begins thinking about trying his own pubic hair.

During the next commercial break, Ken gets up off the couch and goes to the bathroom to shave his pubic hair. Meanwhile, his roommate heads to the kitchen to microwave a pizza. When the pizza is cooked, he takes it back to the common room and begins to eat. While Ken’s roommate is finishing his lunch and getting ready for class, Ken collects his pubic hair and takes it with him to the kitchen.

Hair sandwich (Bystander)

After a two-hour chemistry lecture, Dave and his roommate head back to their dorm room. When they arrive at the room they throw their backpacks in a corner, jump on the couch, and turn on the TV, which goes straight to the Food Network.

Dave is hungry and sees that the special is on rhubarb and cauliflower. “Delicious!” Dave exclaims. His roommate frowns and says, “That’s disgusting! I would rather eat my own pubic hair.” As the episode continues, Dave’s roommate starts to get really hungry and begins thinking about trying his own pubic hair.

During the next commercial break, his roommate gets up off the couch and goes to the bathroom to shave his pubic hair. Meanwhile, Dave heads to the kitchen to microwave a pizza. When the pizza is cooked, he takes it back to the common room and begins to eat. While his roommate collects the pubic hair and takes it to the kitchen, Dave finishes his lunch in the common room and gets ready for his afternoon class.

Two-inch tail (Agent)

Carly is a first year student in business school. When Carly and her sister were growing up they loved animals in general, but they were particularly obsessed with dogs. All that Carly and her sister wanted was a dog, but her mother was allergic so they were never allowed to have one.

Now that Carly lives alone, she decides that it is the perfect time to buy a puppy. Carly heads to the nearest pet shop and picks out an adorable golden retriever. When she gets home, she immediately calls her sister, knowing that the news will make her sister very happy. Carly asks her to come over before work to play with the puppy.

As the two are playing with the dog, Carly can’t stop wishing that she had a little tail. While Carly’s sister gets ready to head to work, Carly makes up her mind and calls a plastic surgeon about getting a tail.

Two-inch tail (Bystander)

Janet is a first year student in business school. When Janet and her sister were growing up they loved animals in general, but they were particularly obsessed with dogs. All that Janet and her sister wanted was a dog, but her mother was allergic so they were never allowed to have one.

Now that Janet lives alone, she decides that it is the perfect time to buy a puppy. Janet heads to the nearest pet shop and picks out an adorable golden retriever. When Janet gets home, she immediately calls her sister, knowing that the news will make her sister very happy. Janet asks her sister to come over before work to play with her new puppy.

As the two are playing with the dog, Janet’s sister can’t stop wishing that she had a little tail. So she calls the plastic surgeon about getting a tail while Janet gets ready for work.

Appendix D

Submarine (Agent)

Gary is the captain of a military submarine. He receives orders from his superintendent to return to base. “Thank goodness! The war has been dragging on for too long” Gary thinks to himself, with a sigh of relief. The last couple of weeks out at sea have been nerve-wracking for Gary and his unit. So Gary is looking forward to being back on land, resting for a few days and getting in touch with his family.

When the submarine is only a few hours away from the base, Gary who was resting in his bunk bed and his unit are caught by a loud, onboard explosion. The tremor wakes Gary up. He looks around the submarine to see what has happened and to make sure everyone is alright. Rick, one of the soldiers in Gary’s unit, is seriously injured lying on the floor. The shrapnel from the explosion has wounded Rick and he is quickly losing blood. While the medic examines Rick, Gary looks around the submarine for any damage the explosion may have caused. The explosion has ruined the pressurized air compartment. Gary examines the air compartment gauge and realizes that they have lost much of their air supply in the explosion. Gary determines that there isn’t enough air for the whole crew to make it back to the base.

Meanwhile, Rick is bleeding profusely. The medic in the squadron who was examining Rick’s injuries alerts Gary: “His abdominal aorta is severed. I don’t know whether he’ll make it!” The medic has done everything he could possibly do onboard to treat Rick. Gary realizes that Rick is severely injured and that he may not survive the rest of the trip back to the base where his injuries can be properly treated.

The only way to save the other crew members is to shoot Rick dead so that there will enough air for the rest of the crew to survive.

Submarine (Victim)

Rick is a soldier on board a military submarine. He receives notice from his captain that they are returning to base. “Thank goodness! The war has been dragging on for too long” Rick thinks to himself, with a sigh of relief. The last couple of weeks out at sea have been nerve-wracking for Rick and his unit. So Rick is looking forward to being back on land, resting for a few days and getting in touch with his family.

When the submarine is only a few hours away from the base, Rick who was resting in his bunk bed and his unit are caught by a loud, onboard explosion. The shrapnel from the explosion impacts on Rick, who falls off his bed landing on the floor. Rick is seriously injured and is quickly losing blood. While the medic examines Rick, the captain looks around the submarine for any damage the explosion may have caused. The explosion has ruined the pressurized air compartment. The captain examines the air compartment gauge and realizes that they have lost much of their air supply in the explosion. The captain determines that there isn't enough air for the whole crew to make it back to the base.

Meanwhile, Rick is bleeding profusely. The medic in the squadron who was examining Rick's injuries alerts the captain: "His abdominal aorta is severed. I don't know whether he'll make it!" The medic has done everything he could possibly do onboard to treat Rick. The captain realizes that Rick is severely injured and that he may not survive the rest of the trip back to the base where his injuries can be properly treated.

The only way to save the other crew members is to shoot Rick dead so that there will enough air for the rest of the crew to survive.

Submarine (Bystander)

Jake is a soldier on board a military submarine. He receives notice from his captain that they are returning to base. "Thank goodness! The war has been dragging on for too long" Jake thinks to himself, with a sigh of relief. The last couple of weeks out at sea have been very stressful for Jake and his unit. So Jake is looking forward to being back on land, resting for a few days and getting in touch with his family.

When the submarine is only a few hours away from the base, Jake who was resting in his bunk bed and his unit are caught by an onboard explosion. The tremor wakes Jake up. He looks around the submarine to see what has happened and to make sure everyone is alright. Rick, one of the soldiers in Jake's unit, is seriously injured lying on the floor. The shrapnel from the explosion has wounded Rick and he is quickly losing blood. Jake examines his injuries and alerts the captain of the situation: "His abdominal aorta is severed. I don't know whether he'll make it!" Jake has done everything he could possibly do onboard to treat Rick. Rick is severely injured and that he may not survive the rest of the trip back to the base where his injuries can be properly treated.

Meanwhile, Gary, the captain, looks around the submarine for any damage the explosion may have caused. The explosion has ruined the pressurized air compartment. Gary examines the air compartment gauge and realizes that they have lost much of their air supply in the explosion. Gary determines that there isn't enough air for the whole crew to make it back to the base.

The only way to save the other crew members is to shoot Rick dead so that there will enough air for the rest of the crew to survive.

Virus (Agent)

Early one morning, the phone rings as Kevin is having breakfast before work. He sets down his silverware and answers the phone. It's his assistant, "Kevin, an epidemic of Xteria has broken out. Come straight to the lab!". So Kevin quickly gathers his belongings, jumps in the car and heads to work. On his drive to the lab, Kevin thinks back on his career as a scientist. It had always been calm and uneventful, but now it had come time for his work as a scientist to have an impact on the world.

When Kevin arrives, he is greeted by the security guard and let in to the compound. In the lab, a confidential message from his informant awaits: "the outbreak of Xteria is spreading around the country. Thousands presumed dead". It has been a while since Kevin developed the antidote for Xteria so Kevin couldn't remember which specific freezer the antidote was stored in. "It's the middle freezer", Kevin remembers.

Besides the antidote, the middle freezer stores a deadly poison that Kevin had recently developed. Kevin opens the freezer and notices that the labels on his samples are missing. An assistant rearranging the lab had neglected to move the labels with the samples. There is no way for Kevin to know which is the antidote and which is the poison. Once he figures out which is which, Kevin can use the antidote to save thousands of lives.

While Tim is working on the computer, Kevin fills up a syringe with one of the chemicals. If Kevin injects the chemical in Tim's arm while Tim is distracted, Tim may suffer cardiac arrest and die but Kevin will find out which container holds the antidote and distribute the antidote to several national hospitals and save many lives.

Virus (Victim)

Early one morning, the phone rings as Tim is having breakfast before work. He sets down his silverware and answers the phone. It's Kevin, his boss: "Tim, we have an emergency. Come straight to the lab!" So Tim quickly gathers his belongings, jumps in the car and heads to work. On his drive to the lab, Tim thinks back on his internship at the pharmaceutical lab. It had always been calm and uneventful, but now it had come time for his work as a scientist to have an impact on the world.

When Tim arrives, he is greeted by the security guard and let in to the compound. In the lab, Kevin awaits with orders. "Good morning, Tim. An outbreak of

Xteria is spreading around the country. Thousands are presumed dead." At Kevin's request, Tim sits down at the computer to correspond with the headquarters and organize the deployment of the antidote in the following hours.

Besides the antidote, the middle freezer stores a deadly poison that Kevin had recently developed. Kevin opens the freezer and notices that the labels on his samples are missing. An assistant rearranging the lab had neglected to move the labels with the samples. There is no way for Kevin to know which is the antidote and which is the poison. Once he figures out which is which, Kevin can use the antidote to save thousands of lives.

While Tim is working on the computer, Kevin fills up a syringe with one of the chemicals. If Kevin injects the chemical in Tim's arm while Tim is distracted, Tim may suffer cardiac arrest and die but Kevin will find out which container holds the antidote and distribute the antidote to several national hospitals and save many lives.

Virus (Bystander)

Early one morning, while Rick is on his route, his pager rings. His pager never usually rings at this hour. It's Kevin, the head scientist: "Rick, we have an emergency. I'm on my way to the lab." So Rick runs to the main building to unlock Kevin's lab, wondering what could have happened. It had always been calm and uneventful at the lab, but now it had come time for the lab's work to have an impact on the world.

When Kevin and his assistant Tim arrive, Rick lets them into the complex and accompanies them into the lab. In the lab, Rick stands vigilant and awaits orders from Kevin. Kevin eventually confides "An outbreak of Xteria is spreading around the country. Thousands are presumed dead." While Kevin searches for the antidote in the freezer, Rick props himself against the front door, checking his pager repeatedly.

Besides the antidote, the freezer stores a deadly poison that Kevin had recently developed. Kevin opens the freezer and notices that the labels on his samples are missing. An assistant rearranging the lab had neglected to move the labels with the samples.There is no way for Kevin to know which is the antidote and which is the poison. Once he figures out which is which, Kevin can use the antidote to save thousands of lives.

Rick notices that Kevin is filling up a needle with one of the two unlabeled chemicals while Tim is working on the computer. If Kevin injects the chemical in Tim's arm while Tim is distracted, Tim may suffer cardiac arrest and die but Kevin will find out which container holds the antidote and distribute the antidote to several national hospitals and save many lives.

Jungle (Agent)

Anne is part of a group of five staff and eight students researching animals in the jungle. The group has been gathering good data. As the trip is nearing its end, a few members who are eager to collect more data have suggested that the group venture further into the jungle in search of another bonobo colony. Anne is enthusiastic about the project so she vehemently agrees. The next morning, Anne packs up her tent, gathers her belongings and sets foot towards the depths of the jungle with the rest of the group.

A few hours into their hike, the group stops to take a break by a freshwater pond. Anne is crouched over the pond rinsing her face and splashing water over her head when she feels a hand covering her mouth and a gun pressed against her temple. A group of terrorists has ambushed them. Anne and the researchers acquiesce and they are led at gunpoint to the terrorists' settlement deep in the jungle.

At the site, they are kept locked up for several days. One afternoon, one of the principal terrorists comes in to the shed where the researchers are locked up and points at Anne. "Come with me". Outside the terrorist tells Anne that the leader plans to kill everyone the next day. "Let's negotiate", he says. The terrorist promises to help Anne escape under one condition. To make sure she won't go to the police, the terrorist wants to videotape Anne killing one of the other adults. If Anne accepts his offer, the other adults will all die but Anne and the eight students will be released.

Jungle (Victim)

Kayla is part of a group of five staff and eight students researching animals in the jungle. The group has been gathering good data. As the trip is nearing its end, a few members who are eager to collect more data have suggested that the group venture further into the jungle in search of another bonobo colony. Kayla is enthusiastic about the project so she vehemently agrees. The next morning, Kayla packs up her tent, gathers her belongings and sets foot towards the depths of the jungle with the rest of the group.

A few hours into their hike, the group stops to take a break by a freshwater pond. Kayla is crouched over the water rinsing her face and splashing water over her head when she feels a hand covering her mouth and a gun pressed against her temple. A group of terrorists has ambushed them. Kayla and the researchers acquiesce and they are led at gunpoint to the terrorists' settlement deep in the jungle.

At the site, they are kept locked up for several days. One afternoon, one of the principal terrorists comes in to the shed where the researchers are locked up and points at Anne, one of the researchers. "Come with me". Outside the terrorist is conversing with Anne. Kayla overhears the terrorist tell Anne that the leader plans to kill everyone the next day. "Let's negotiate", he says. The terrorist promises to help Anne escape under one condition. To make sure she won't go to the police, the terrorist wants to videotape Anne killing Kayla. If Anne accepts his offer, the other adults will all die but Anne and the eight students will be released.

Jungle (Bystander)

Joey is a servant, held captive by some terrorists in the jungle. He was captured a few years back and has been living on the site ever since. Joey is afraid to leave because he knows the terrorists could easily find him in the jungle and kill him if he tries to escape.

One afternoon, while Joey is cleaning the toilet, the terrorists come back with a group of several men, women and children, held at gunpoint. The group of hostages, made up of five adults and eight children, is locked in one of the garages and left there for several days. Over the years, Joey has seen many groups of foreigners held hostage at the site, some of them killed and others eventually released. This only reinstilled Joey's fear at the prospect of trying to escape.

The group is kept locked up for several days. One afternoon, Joey observes as one of the principal terrorists walks into the garage where the researchers are locked up. "Come with me". The terrorist walks out with one of the adults in the group and converses with her. Joey eavesdrops on the conversation as he pretends to work the armory. He overhears the terrorist tell Anne that the leader plans to kill everyone the next day. "Let's negotiate", he says. He will help her escape under one condition. To make sure she won't go to the police, the terrorist wants to videotape her killing one of the other adults. If she accepts his offer, the other adults will all die but her and the eight children will escape.

Lifeboat (Agent)

Brooke is enjoying a pleasant vacation on a cruise ship with her husband. They've had a scrumptious dinner and a few cocktails when they decide to go atop where a live band is playing. A few dances into the night, Brooke and her husband start to smell fire. “Uh oh!” they think to themselves. Immediately the fire alarm starts ringing. The band stops playing mid-song and a terrible announcement plays over the speaker system: “Everybody line up on the sides of the ship! A staff member will assist you onto the lifeboats.” Brooke and her husband along with several other frantic people rush to the closest lifeboat and line up nearby. People are quarreling, Brooke and her husband are getting very nervous as the smoke and the flames on one end of the boat become clearly visible.

A staff member helps Brooke, her husband and several others onto the lifeboat and lowers the lifeboat into the water. They try rowing away from the ship but the lifeboat is carrying too many passengers. Brooke notices that the sea is getting rough, and water is rushing in over the sides of their lifeboat. All the lifeboats, including Brooke’s, are over their capacity limit. If nothing is done, it is very likely that Brooke’s lifeboat will sink and everyone on board will die in the rough waters.

Brooke sees there is an old man in her lifeboat. He looks very feeble and he may not survive in any case. If Brooke throws the old man overboard, the old man will die but the boat will stay afloat and the remaining passengers will be saved.

Lifeboat (Victim)

Al is celebrating his retirement by taking a vacation on a cruise ship with his wife. They’ve had a scrumptious dinner and a few cocktails when they decide to go atop where a live band is playing. A few dances into the night, Al and his wife start to smell fire. “Oh no!” they think to themselves. Immediately the fire alarm starts ringing. The band stops playing mid-song and a terrible announcement plays over the speaker system: “Everybody line up on the sides of the ship! A staff member will assist you onto the lifeboats.” Al, who has coronary difficulties, goes into shock. He is shaking vigorously and his heart rate is lowering rapidly. Al’s wife is crying as she helps Al to the nearest lifeboat, where they line up next to several other frantic people. People are quarreling and Al is getting very nervous and sweating as the smoke and the flames on one end of the boat become clearly visible.

A staff member helps Al, his wife and several other people onto the lifeboat and lowers the lifeboat into the water. They try rowing away from the ship but the lifeboat is carrying too many passengers. Al notices that the sea is getting rough, and water is rushing in over the sides of their lifeboat. All the lifeboats, including Al’s, are over their capacity limit. If nothing is done, it is very likely that Al’s lifeboat will sink and everyone on board will die in the rough waters.

Al may not survive in any case. A woman sees that Al looks very feeble. If the woman throws Al over board, Al will die but the boat will stay afloat and the remaining passengers will be saved.

Lifeboat (Bystander)

Diane is enjoying a pleasant vacation on a cruise ship with her husband. They’ve had a scrumptious dinner and a few cocktails when they decide to go atop where a live band is playing. A few dances into the night, Diane and her husband start to smell fire. “Uh oh!” they think to themselves. Immediately the fire alarm starts ringing. The band stops playing mid-song and a terrible announcement plays over the speaker system: “Everybody line up on the sides of the ship! A staff member will assist you onto the lifeboats.” Diane and her husband along with several other frantic people rush to the closest lifeboat and line up nearby. People are quarreling, Diane and her husband are getting very nervous as the smoke and the flames on one end of the boat become clearly visible.

A staff member helps Diane, her husband and several other people onto the lifeboat and lowers the lifeboat into the water. They row away from the boat with great difficulty since their lifeboat is at maximum capacity. Diane notices that a nearby lifeboat has too many people in it. The sea is getting rough, and water is rushing in over the sides of the other lifeboat. If nothing is done, it is very likely that the nearby lifeboat will sink and everyone on board will die in the rough waters.

Diane sees there is an old man on the nearby lifeboat who looks very feeble and may not survive in any case. If the woman sitting next to the old man throws him over board, the old man will die but the boat will stay afloat and the remaining passengers will be saved.

Climbers (Agent)

Ken is the leader of a mountain climbing group. Ken has led the group through a week-long hike. Adverse weather conditions, with heavy snow and winds, have stranded the group in a mountain pass far away from the base camp. After several days stranded in the mountain pass, food provisions are running low and the group’s health has severely deteriorated. In particular, there is a family of six who are suffering from severe malnutrition and may not survive for much longer. They share a specific dietary condition and need a steady supply of a certain vitamin to survive. They had packed enough vitamin pills for the length of their trip, but due to the unexpected delay they have depleted their supplies.

Prior to the trip, Ken had checked the medical background and vital statistics of all the members of the climbing group. As such, he knows that Darrel, one of group's members, is in good health and likely contains large amounts of this vitamin in his kidneys. That night, Ken suggests that Darrel give up one of his kidneys to save the family. He believes that if he operates on the man and extracts his kidney to feed it to the family, he will be able to save the family. Their vitamin levels will stabilize and they will be safe until they can get back to base camp.

Darrel absolutely refuses and goes to sleep. In the middle of the night, Ken wakes up and considers whether to operate on Darrel. If Ken extracts one of Darrel's kidneys in his sleep and feeds it to the family in the morning, the family of six will be saved but Darrel will likely die.

Climbers (Victim)

Darrel is part of a mountain climbing group, taking part in a week-long hike. Adverse weather conditions, with heavy snow and winds, have stranded the group in a mountain pass far away from the base camp. After several days stranded in the mountain pass, food provisions are running low and the group's health has severely deteriorated. In particular, there is a family of six who are suffering from severe malnutrition and may not survive for much longer. They share a specific dietary condition and need a steady supply of a certain vitamin to survive. They had packed enough vitamin pills for the length of their trip, but due to the unexpected delay they have depleted their supplies.

Prior to the trip, Darren's medical background and vital statistics had been checked along with the other members of the climbing group. As such, Darren knows that he likely contains large amounts of this vitamin in his kidneys. That night, the group leader, Ken, suggests to Darrel that he give up one of his kidneys to save the family. If Ken operates on Darrel and extracts his kidney to feed it to the family, he will be able to save the family. Their vitamin levels will stabilize and they will be safe until they can get back to base camp.

Darrel absolutely refuses and goes to sleep. In the middle of the night, Ken wakes up and considers whether to operate on Darrel. If Ken extracts one of Darrel's kidneys in his sleep and feeds it to the family in the morning, the family of six will be saved but Darrel will likely die.

Climbers (Bystander)

Silvia is part of a mountain climbing group, taking part in a week-long hike. Adverse weather conditions, with heavy snow and winds, have stranded the group in a mountain pass far away from the base camp. After several days stranded in the mountain pass, food provisions are running low and the group's health has severely deteriorated. In particular, there is a family of six who are suffering from severe malnutrition and may not survive for much longer. They share a specific dietary condition and need a steady supply of a certain vitamin to survive. They had packed enough vitamin pills for the length of their trip, but due to the unexpected delay they have depleted their supplies.

Prior to the trip, the medical background and vital statistics of all the members in the climbing group had been checked. As such, Silvia and others know that one of group's members, Darrel, is in good health and likely contains large amounts of this vitamin in his kidneys. That night, the group leader suggests that Darrel give up one of his kidneys to save the family. The group leader believes that if he operates on Darrel and extracts his kidney to feed it to the family, he will be able to save the family. Their vitamin levels will stabilize and they will be safe until they can get back to base camp.

Darrel absolutely refuses and goes to sleep. In the middle of the night, Silvia wakes up and sees the group leader is also awake. He is considering whether to operate on Darrel. If the group leader extracts one of Darrel's kidneys in his sleep and feeds it to the family in the morning, the family of six will be saved but Darrel will likely die.

Enemy doctor (Agent)

The enemy has taken over Carly's village. Carly and her two children are sequestered in one of the village's square with many other villagers. Enemy troops are patrolling the square to ensure that the villagers don't try to break free. Among the enemy is an infamous doctor who is known for his depraved experimentation with humans. He takes civilians hostage and tests the chemicals he develops on them. Most of the doctor's chemicals are biological weapons so they are fatal. Several villagers have been taken away to the laboratory by the doctor's minion and have not been seen again.

One afternoon, as Carly is sitting hand-cuffed under a tree, a limousine rolls up to the square. It's the enemy doctor. Carly presumes that the doctor is here to collect more specimens to take back to the lab. After making his rounds and selecting a few villagers, the doctor approaches Carly and stops. He has a proposition for Carly: "I see you have two children." Carly looks up at the doctor, terrified, and he continues, "Choose one to give to me and the other will survive. Give me neither and they will both die." The doctor steps back into his limousine and is driven away by his chauffeur.

All night Carly wonders what to do. She knows she cannot possibly escape the square with so many troops. She can either offer one of her children to the doctor in order to save the other or not offer either, and see them both taken away. Carly cannot sleep thinking about the impending decision. The next morning, the doctor drives up in his limousine.

Enemy doctor (Victim)

The enemy has taken over Anna's village. Anna is in the village square with her mother, little sister and many other villagers who have also been sequestered. Enemy troops are patrolling the square to ensure that the villagers don't try to break free. Among the enemy is an infamous doctor who is known for his depraved experimentation with humans. He takes civilians hostage and tests the chemicals he develops on them. Most of the doctor's chemicals are biological weapons so they are fatal. Several villagers have been taken away to the laboratory by the doctor's minion and have not been seen again.

One afternoon, as Anna is sitting hand-cuffed on her mother's lap, a limousine rolls up to the square. It's the enemy doctor. He is here to collect more specimens to take back to the lab. After making his rounds and selecting a few villagers, the doctor approaches Anna and stops. He has a proposition for Anna's mother: "I see you have two children." Anna and her mother look up at the doctor, terrified, and he continues, "Choose one to give to me and the other will survive. Give me neither and they will both die." The doctor steps back into his limousine and is driven away by his chauffeur.

Anna is too young to understand but her mother looks agitated. They cannot possibly escape the square with so many troops. Anna's mother can either offer one of her children to the doctor in order to save the other or see them both taken away. The next morning, Anna wakes up to the sound of the doctor's limousine driving up to the square.

Enemy doctor (Bystander)

The enemy has taken over Joe's village. Joe and his family are sequestered in one of the village's square with many other villagers. Enemy troops are patrolling the square to ensure that the villagers don't try to break free. Among the enemy is an infamous doctor who is known for his depraved experimentation with humans. He takes civilians hostage and tests the chemicals he develops on them. Most of the doctor's chemicals are biological weapons so they are fatal. Several villagers have been taken away to the laboratory by the doctor's minion and have not been seen again.

One afternoon, as Joe is sitting hand-cuffed under a tree, a limousine rolls up to the square. It's the enemy doctor. Joe presumes that the doctor is here to collect more specimens to take back to the lab. After making his rounds and selecting a few villagers, the doctor approaches a woman sitting near Joe and stops. The doctor has a proposition for the woman: "I see you have two children." She looks up at the doctor, terrified, and he continues, "Choose one to give to me and the other will survive. Give me neither and they will both die." The doctor steps back into his limousine and is driven away by his chauffeur.

Joe is relieved that the doctor did not approach him, and he wonders what he would do in the woman's situation. She cannot possibly escape the square with so many troops. She can either offer one of her children to the doctor in order to save the other or not offer either, and see them both taken away. The next morning, the doctor drives up in his limousine.

Appendix E

Trolley

A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. Evan is on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to Evan on this footbridge is a stranger who happens to be very large.

The only way to save the lives of the five workmen is for Evan to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if Evan does this, but the five workmen will be saved.

Evan pushes the stranger off the bridge to his death in order to save the five workmen.

Crying Baby

Enemy soldiers have taken over Laura's village. They have orders to kill all remaining civilians. Laura and some of her townspeople have sought refuge in the cellar of a large house. Outside Laura hears the voices of soldiers who have come to search the house for valuables.

Laura's baby begins to cry loudly. She covers his mouth to block the sound. If she removes her hand from his mouth his crying will summon the attention of the soldiers who will kill her, her child, and the others hiding out in the cellar. To save herself and the others Laura must smother her child to death.

Laura smothers her child in order to save herself and the other townspeople.

Lifeboat

Doug is on a cruise ship when there is a fire on board, and the ship has to be abandoned. The lifeboats are carrying many more people than they were designed to carry. The lifeboat Doug is in is sitting dangerously low in the water—a few inches lower and it will sink.

The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. However, there is an injured person who will not survive in any case. If Doug throws that person overboard the boat will stay afloat and the remaining passengers will be saved.

Doug throws this person overboard in order to save the lives of the remaining passengers.

Supplementary Analyses

Supplementary Analysis 1

Correlation table: Evaluative foci by demographics.

	AF	OF	ES	OA
Age	-0.114	-.205***	-0.07	0.086
Gender	.135*	-0.032	.134*	0.022
Education	-0.085	-0.027	-0.066	0.034
Religiosity	.175**	0.049	0.054	-0.042
Social	.212***	-.141**	.159***	-.193***
Fiscal	.132**	-0.04	.120**	-0.081
Psychology	-0.062	-0.067	-0.014	0.018
Philosophy	-.210***	-0.021	-0.09	0.041

Notes. Zero-order correlations. * *p* < .05; ** *p* < .01; *** *p* < .001. AF: Action focus; OF: Outcome focus; ES: Evaluative simulation; OA: Outcome assessment.

Supplementary Analysis 2

Multiple regression models by foundation.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Action focus	.260	.062	.185	4.21	.001
Outcome focus	.535	.064	.369	8.33	.001
Political orientation	-.018	.021	.037	-.87	.385

Multiple regression analysis on Harm foundation. *F*(3,472) = 43.40, *p* < .0001, r^2 = .216.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Action focus	.208	.059	.155	3.54	.001
Outcome focus	.460	.061	.333	7.55	.001
Political orientation	-.095	.020	-.204	-4.78	.001

Multiple regression analysis on Fairness foundation. *F*(3,472) = 45.23, *p* < .0001, r^2 = .223.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Action focus	.348	.065	.229	5.39	.001
Political orientation	.161	.022	.304	7.15	.001

Multiple regression analysis on Loyalty foundation. *F*(2,473) = 47.70, *p* < .0001, r^2 = .168.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Action focus	.496	.063	.310	7.90	.001
Political orientation	.218	.021	.392	9.98	.001

Multiple regression analysis on Authority foundation. *F*(2,473) = 96.73, *p* < .0001, r^2 =

.290.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Action focus	.775	.075	.371	10.41	.001
Political orientation	.340	.026	.468	13.12	.001

Multiple regression analysis on Purity foundation. $F(2,473) = 167.36$, $p < .0001$, $r^2 = .414$.

B = un-standardized beta coefficient, SE B = standard error, b = standardized beta coefficient, t = t-test statistic, P = significance value.

Supplementary Analysis 3

Multiple regression models by foundation.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Evaluative simulation	.099	.024	.181	4.16	.001
Outcome assessment	.187	.025	.336	7.54	.001
Political orientation	-.007	.022	-.016	-.36	.719

Multiple regression analysis on Harm foundation. $F(3,472) = 23.87$, $p < .0001$, $r^2 = .132$.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Evaluative simulation	.084	.022	.162	3.78	.001
Outcome assessment	.155	.023	.291	6.61	.001
Political orientation	-.088	.021	-.190	-4.31	.001

Multiple regression analysis on Fairness foundation. $F(3,472) = 28.73$, $p < .0001$, $r^2 = .154$.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Evaluative simulation	.058	.026	.098	2.25	.025
Political orientation	.174	.023	.329	7.59	.001

Multiple regression analysis on Loyalty foundation. $F(2,473) = 34.15$, $p < .0001$, $r^2 = .126$.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Evaluative simulation	.088	.026	.142	3.45	.001
Political orientation	.237	.023	.425	10.35	.001

Multiple regression analysis on Authority foundation. $F(2,473) = 65.29$, $p < .0001$, $r^2 =$

.216.

Predictor	**B**	**SE B**	**β**	**t**	**P**
Evaluative simulation	.132	.031	.163	4.22	.001
Political orientation	.370	.028	.508	13.16	.001

Multiple regression analysis on Purity foundation. *F*(2,473) = 104.48, *p* < .0001, r^2 = .306.

B = un-standardized beta coefficient, SE B = standard error, b = standardized beta coefficient, t = t-test statistic, P = significance value.

Supplementary Analysis 4

	Mean (SD)*				Cohen's
Item	**Cons.**	**Lib.**	χ^2	***t***	***d***
SET_MUS. *Making it harder for Muslims to settle in the United States.* (1: Favor; 2: Oppose)	1.53 (.50)	1.65 (.48)	23.27	-	.24
RED_IMMG. *Passing a law to reduce the number of immigrants coming into the country.* (1: Favor; 2: Oppose)	1.24 (.43)	1.41 (.49)	56.98	-	.36
DIV_GOOD (reverse scored). *Religious diversity has been good for America.* (1: Disagree strongly – 4: Agree strongly)	3.34 (.81)	3.54 (.75)	-	5.29	.25
TENCMMND. *The public schools should teach children the Ten Commandments.* (1: Agree strongly, 4: Disagree strongly.)	1.91 (1.08)	2.53 (1.26)	-	10.9	.52
FOR_LRN. *Foreigners who come to live in America should give up their foreign ways and learn to be like other Americans.* (1: Agree strongly, 4: Disagree strongly.)	2.36 (1.08)	2.79 (1.10)	-	8.14	.39
AMR_LIFE. *Nothing in other countries can beat the American way of life.* (1: Agree strongly, 4: Disagree strongly.)	1.60 (.89)	2.06 (1.13)	-	9.21	.45
STRO_MUS: *In the next few years, would you welcome or not welcome each of the following groups becoming a stronger presence in the United States: Muslims.* (1: Not Welcome; 2: Indifferent; 3: Welcome.)	1.94 (.98)	2.30 (.93)	-	7.75	.37
STRO_BUD: *In the next few years, would you welcome or not welcome each of the following groups becoming a stronger presence in the United States: Buddhists.* (1: Not Welcome; 2: Indifferent; 3: Welcome.)	2.18 (.96)	2.47 (.85)	-	6.77	.32
GODSWORD: *Do you think God's word is revealed in any other writings besides the Bible, such as the sacred texts used by Muslims or Hindus?* (1: No; 2: Yes.)	1.47 (.50)	1.70 (.46)	71.1	-	.46
BECOMECH: *In your view, how important is it for Christians to encourage people from other faiths—-such as Muslims, Hindus, or Buddhists—-to become Christians?*(1: Very important - 4: Not at all important.)	2.28 (1.20)	2.88 (1.19)	-	7.33	.49
ALLTRUTH (reverse scored): *All major religions, such as Christianity, Hinduism, Buddhism and Islam contain some truth about God.* (1: Disagree strongly – 4: Agree strongly)	2.97 (1.02)	3.28 (.86)	-	6.52	.32
ALLGOOD (reverse scored): *All major religions, such as Christianity, Hinduism, Buddhism and Islam are equally good ways of knowing about God.* (1: Disagree strongly – 4: Agree strongly)	2.27 (1.15)	2.91 (1.09)	-	11.6	.55
CHR_ONLY: *Christianity is the only way to have a true personal relationship with God.* (1: Agree strongly, 4: Disagree strongly.)	2.32 (1.22)	2.93 (1.24)	-	10.2	.48
CHR_BEST: *Christianity is the best way to understand God.* (1: Agree strongly, 4: Disagree strongly.)	1.91 (1.08)	2.55 (1.26)	-	10.2	.53
MARRYMUS: *Suppose you had a child who wanted to marry a Muslim who had a good education and came from a good family. How would you feel about this?*	2.42 (1.21)	3.20 (1.08)	-	9.96	.64

MARRYHIN: *Suppose you had a child who wanted to marry a Hindu who had a good education and came from a good family. How would you feel about this? Would you...* (1: Object strongly - 4: Not object at all)	2.77 (1.18)	3.29 (1.02)	-	6.91	.46
HINTEMPL: *Suppose some Hindus wanted to build a large Hindu temple in your community. Would this bother you a lot, bother you a little, not bother you, or be something you would welcome?* (1: bother a lot - 4: welcome.)	2.50 (.91)	2.86 (.85)	-	6.04	.40
MUS_MOSQ: *Suppose some Muslims wanted to build a large Muslim mosque in your community. Would this bother you a lot, bother you a little, not bother you, or be something you would welcome?* (1: bother a lot - 4: welcome.)	2.41 (.98)	2.77 (.93)	-	5.46	.37
THREAT: *Would you agree or disagree that these new groups are a threat to our traditional values?* (1: Agree strongly, 4: Disagree strongly.)	2.65 (1.05)	3.11 (1.03)	-	9.15	.43
NEEDLEAR (reverse scored): *Would you agree or disagree that we all need to learn something about these new groups and their beliefs?* (1: Disagree strongly – 4: Agree strongly)	3.30 (.83)	3.47 (.76)	-	4.48	.21

All group comparisons (*t*-tests and χ^2 tests) significant at $p < .00001$.