

Component Neural Networks of Morality

by

Lawrence Ngo

Department of Neurobiology  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Scott A. Huettel, Supervisor

\_\_\_\_\_  
Walter Sinnott-Armstrong

\_\_\_\_\_  
Michael L. Platt

\_\_\_\_\_  
R. Alison Adcock

\_\_\_\_\_  
J.H. Pate Skene

Dissertation submitted in partial fulfillment of  
the requirements for the degree of Doctor  
of Philosophy in the Department of  
Neurobiology in the Graduate School  
of Duke University

2014

ABSTRACT

Component Neural Networks of Morality

by

Lawrence Ngo

Department of Neurobiology  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Scott A. Huettel, Supervisor

\_\_\_\_\_  
Walter Sinnott-Armstrong

\_\_\_\_\_  
Michael L. Platt

\_\_\_\_\_  
R. Alison Adcock

\_\_\_\_\_  
J.H. Pate Skene

An abstract of a dissertation submitted in partial  
fulfillment of the requirements for the degree  
of Doctor of Philosophy in the Department of  
Neurobiology in the Graduate School of  
Duke University

2014

Copyright by  
Lawrence Ngo  
2014

## **Abstract**

Moral cognition represents a foundational faculty of the human species. Our sense of morality develops beginning at a very young age, and its dysfunction can lead to devastating mental disorders. Given its central importance, it has fittingly garnered the attention of thinkers throughout the ages. For millennia, philosophers have pondered what it is to be right or wrong, good or bad, virtuous or vicious. For centuries, psychologists have elucidated how people acquire and act upon a sense of morality. More recently in the last decade, neuroscientists have embarked on a project to study how morality arises from computations in the brain. However, this latest project has been fragmented: researchers have largely studied various neural components of morality – including emotion, value, and mentalizing – in isolation. This has resulted in an informal and disjointed model for the neural mechanisms of morality. This dissertation is concerned with more formally identifying neural components and their influences on each other in the context of moral cognition.

In Chapter 2, I study how the component neural networks of moral cognition may be involved in distinct aspects of a single decision by employing a complex clinical decision making task involving the disclosure of conflicts of interest. I show that for a given decision, the magnitude of conflict of interest is tracked by mentalizing networks, while the degree of disclosure-induced behavioral change exhibited by participants is

predicted by value networks. In Chapter 3, I move beyond the informal model of morality used in Chapter 2 and previous literature by devising a methodology to identify hierarchical ontologies of neural circuits; such an approach can have implications on further discussions of morality, and more generally, on other aspects of cognitive neuroscience. From this, I present the 50 elemental neural circuits that are fundamental to human cognition and explore how these elements can differentially combine to form emergent neural circuits. In Chapter 4, I use these advances to address morality, uncovering its relevant component neural networks in a data-driven way. I show that neural circuits important in supporting higher-level moral computations include mentalizing and taste. In Chapter 5, I demonstrate an important complexity in a compositional model of morality. I show that one of the components of moral cognition, mentalizing, can paradoxically be influenced by moral judgments themselves. To conclude, I highlight the implications of both theoretical and methodological advances. The hierarchical ontologies of neural circuits may be a profitable framework for the future characterization and study of mental disorders; and to effectively study these circuits, the use of moral judgment and decision-making paradigms will be effective experimental tasks, considering the centrality of moral cognition to who we are, whether in health or illness.

# Dedication

For Jen.

# Contents

Abstract .....	iv
List of Tables .....	xi
List of Figures .....	xii
Acknowledgements .....	xvi
1. Introduction .....	1
1.1 Thinking Hard about Morality for Millennia .....	4
1.1.1 Normative Perspectives of Morality .....	4
1.1.2 Psychological Perspectives of Morality .....	6
1.1.3 Neuroscientific Perspectives of Morality .....	8
1.1.3.1 Moral Emotion .....	8
1.1.3.2 Moral Value .....	13
1.1.3.3 Moral Mentalizing .....	16
1.1.3.4 Moral Component Interactions .....	20
2. Mentalizing and Value in the Disclosure of Conflicts of Interest .....	24
2.1 Introduction .....	24
2.2 Behavioral Methods .....	26
2.2.1 Participants .....	26
2.2.2 fMRI Task .....	26
2.2.3 Analysis .....	29
2.3 Imaging Methods .....	30

2.3.1 Acquisition and Preprocessing.....	30
2.3.2 fMRI Analysis .....	31
2.3.3 Neurosynth Decoding.....	32
2.4 Behavioral Results .....	32
2.5 fMRI Results .....	33
2.5.1 DMPFC tracks the patient outcomes .....	33
2.5.2 Striatal activation predicts individual differences in responsiveness to disclosure policy.....	36
2.6 Discussion.....	39
3. The Neural Elements and Compounds of Cognitive Neuroscience.....	44
3.1 Introduction.....	44
3.2 Materials and Methods .....	45
3.2.1 Neurosynth.....	45
3.2.2 ICA Decomposition.....	47
3.2.3 Descriptive Thresholding.....	48
3.2.4 Clustering .....	49
3.3 Results .....	50
3.3.1 Component Categorization.....	50
3.3.2 The “Periodic Table of Neural Elements” .....	51
3.3.3 Brain Region Parcellation.....	52
3.3.4 Term types.....	54
3.4 Discussion.....	60

3.4.1 Theoretical Implications of Elemental Distinction .....	60
3.4.2 Semantics in Relation to Different Types of Components .....	61
3.4.3 Component Filtering .....	62
3.4.4 Neural Elements and Diversity .....	64
4. The Elemental Composition of Morality as a Neural Compound .....	67
4.1 Introduction .....	67
4.2 Methods .....	69
4.3 Results .....	70
4.4 Discussion .....	73
5. Two Distinct Moral Mechanisms for Ascribing and Denying Intentionality .....	79
5.1 Introduction .....	79
5.2 Methods .....	82
5.2.1 Experiment 1: Campus Behavioral Experiment .....	82
5.2.1.1 Participants .....	82
5.2.1.2 Task .....	82
5.2.1.3 Analysis .....	83
5.2.2 Experiment 2: Online Behavioral Experiment .....	85
5.2.2.1 Participants .....	85
5.2.2.2 Task .....	87
5.2.2.3 Analysis .....	87
5.2.3 Experiment 3: fMRI Experiment .....	88
5.2.3.1 Participants .....	88

5.2.3.2 Stimuli and Tasks.....	88
5.2.3.3 Behavioral Analysis: Mediation Models.....	91
5.2.3.4 fMRI Analysis: Image Acquisition and Pre-processing.....	92
5.2.3.5 fMRI Analysis: Generalized Linear Model.....	92
5.2.3.6 Neural Mediation Model .....	93
5.3 Results .....	95
5.3.1 Experiment 1 .....	95
5.3.2 Experiment 2 .....	98
5.3.3 Experiment 3 .....	101
5.4 Discussion.....	106
6. General Discussion .....	110
6.1 Overview .....	110
6.2 What is the essence of morality? .....	111
6.3 Methodological Advancements.....	115
6.4 Practical Implications.....	116
References .....	124
Biography .....	139

## List of Tables

Table 1: Clusters of activation for the intersection of voxels that significantly tracked the level of conflict of interest for both conflict and non-conflict trials. All indicated clusters passed a voxel significance threshold of $z > 2.3$ and were whole brain corrected at $p < 0.05$ . .....	35
Table 2: Maxima for clusters whose activity significantly tracked the level of patient-biased responses to disclosure policy. All indicated clusters passed a voxel significance threshold of $z > 2.3$ and were whole-brain corrected at $p < 0.05$ . .....	36
Table 3: Top ten Neurosynth terms associated with clusters of interest from Tables S1 and S2. This includes all of the clusters in Table 1 for DMPFC and the ventral striatum cluster from Table 2. ....	38
Table 4: Decoding using the neural elements provides a higher degree of separation between competing reverse inferences in comparison to that shown in Table 3. ....	117
Table 5: Proposed mapping between domains proposed by the NIMH's RDoC and the neural elements described in Chapter 3. ....	122

## List of Figures

Figure 1: Survey of major work in the literature pertaining to moral judgments and decisions across the three domains of emotion, mentalizing, and value..... 23

Figure 2: Task structure. After the experiment, one trial was chosen at random, and the physician's advice will be given to different participant. Both physician and patient are paid according to the option chosen by the patient in virtue of the received advice in the form of a gift card..... 28

Figure 3: Dorsomedial prefrontal cortex tracks the difference in possible patient outcomes. (A) The intersection between significantly active voxels found to significantly track the magnitude of difference between patient payments for conflict and non-conflict trials yielded a cluster of activation in the dorsomedial prefrontal cortex (whole-brain corrected at  $p < 0.05$ ; max voxel, MNI: (4, 54, 26),  $z = 8.37$ ). Heightened activation in this region was associated with larger differences between outcomes between Treatment A and B for the patient. (B) Parameter estimates drawn from the region shown in (A) are significantly positive for differences in outcomes for the patient. However, those drawn from the parameter estimates of outcomes for self were not significantly different from zero. (n = 30)..... 35

Figure 4: Striatal activity is associated with response to disclosure policy for conflicts of interest. (A) Individual differences in behavioral response to disclosure policy for conflict trials is positively correlated with activity in the striatum (whole-brain corrected at  $p < 0.05$ ; striatal peak, MNI:(-6, 20, 08),  $z = 3.36$ ). (B) The same comparison for non-conflict trials was not significant ( $R = 0.095$ ;  $p = 0.61$ ). The scatter plot for conflict trials recapitulates the data in (A) and is provided for illustrative purposes. The “Disclosure Influence on Advice” metric is a difference of differences measure. First, the difference in accumulated payoffs of given advice for trials between patient and self are calculated separately for disclosure and non-disclosure trials. The difference is then taken between these two conditions. The “Influence of Disclosure in Ventral Striatum” is the difference in parameter estimates across disclosure and non-disclosure trials drawn from the ventral striatum ROI shown in (A). (n = 30)..... 37

Figure 5: The “Periodic Table of Neural Elements.” From left to right, the neural elements are generally arranged by relevance to sensory input and motor output. Color categories were initially derived from the hierarchical clustering algorithm, such that underlined elements (or italicized for the green group which has two clusters) within the same color group had some degree of clustering according to the algorithm. The other

elements belonging to each group were manually arranged into groups for visualization purposes. Numbers indicated for each element denote the numbering of the components from the most to least amount of variance explained from the original ICA decomposition. .... 53

Figure 6: Default mode and mentalizing neural elements have dissociable patterns of activation in temporoparietal junction, medial prefrontal cortex, and precuneus. Images are thresholded at  $z > 3$ . The maps are drawn from the spatial maps resulting from ICA decomposition of terms in the Neurosynth database. .... 54

Figure 7: Elemental terms have a strong one-to-one relevance to a particular neural element. In this case, “auditory” is highly associated with neural element 2, while “reward” strongly corresponds to neural element 10. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle. .... 55

Figure 8: Sub-elemental terms do not have a strong one-to-one mapping to any particular neural element, but rather have a high relevance for multiple neural elements. Here, “motor” and “social” have multiple strong peaks. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle. .... 56

Figure 9: Non-elemental terms do not have any particularly specific correspondence to any single neural elements and do not have moderate to low association across all elements. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle. .... 58

Figure 10: Neural compounds are composed of the combination of several different neural elements. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle. .... 59

Figure 11: Composition of *moral* from the Neurosynth database. The methods used from Chapter 3 were applied, and 15 elements with the greatest relevance to *moral* are listed here in descending order. The labels for these elements are taken from Figure 5. .... 71

Figure 12: *Disgust*, as a common discussed component of morality, is not characterized as being a neural element. Its relevance is widely distributed among multiple neural elements. The methods used from Chapter 3 were applied, and 15 elements with the greatest relevance to *disgust* are listed here in descending order. The labels for these elements are taken from Figure 5. .... 72

Figure 13: Asymmetries in intentionality are robust across three different methods of experimentation. (A) In the fMRI version of the task, participants read and responded to two versions of each general story. These versions differed in whether the agent’s actions lead to morally negative or positive consequences (40 pairs for 80 vignettes total). Participants provided ratings of intentionality on a scale from 1 (completely unintentionally) to 8 (completely intentionally), and the direction of the scale was counterbalanced trial-by-trial. Reported imaging results are derived from data collected during the “Knowledge” epoch. The ITI was 2 s. (B) Participants consistently rated actions in negative conditions as being more intentional than those in positive conditions across three different experiments. Model-free means are presented along with 95% confidence intervals for comparison across three different experimental designs. \*Indicates that the means are different according to paired *t*-tests for experiment 1 and according to hierarchical, mixed-effect modeling for experiments 2 and 3..... 96

Figure 14: Emotional salience does not account for differences in intentionality ratings between outcomes with different emotional valence, but it does predict intentionality for negative outcomes. Participants (n=386) on AMT were presented three versions of scenario #4 differing in valence. (A) All pairwise comparisons among the three conditions were significantly different from one another. Participants ascribed higher intentionality for negative compared to positive (paired  $t(192) = 11.3, p < 0.0001$ ), higher for negative compared to neutral (paired  $t(192) = 8.19, p < 0.0001$ ) and lower intentionality to positive compared to neutral (paired  $t(192) = 2.58, p < 0.01$ ). The data from negative and positive conditions were also presented in Figure 13B. (B) The neutral condition had significantly lower ratings of salience than negative (paired  $t(192) = 18.03, p < 0.0001$ ) and positive conditions (paired  $t(192) = -17.8, p < 0.0001$ ). Negative conditions did have higher salience ratings than those for positive conditions (paired  $t(192) = 2.28, p = 0.02$ ). Error bars indicate 95% confidence interval. (C) For negative conditions, salience ratings were positively correlated with those for intentionality. The same was not found in neutral conditions (D) or in positive conditions (E). Density plots are overlaid with a regression line with 95% confidence interval. \*All pairwise comparisons are significantly different from one another according to a paired *t*-test. .... 99

Figure 15: Converging behavioral and neural evidence suggests that *Ascription* leads to higher intentionality through an emotional mechanism while *Denial* leads to lower intentionality and is dependent on *statistical normativity*. (A) Behaviorally, *emotional reaction* significantly predicts intentionality ratings for negative conditions but not for positive conditions. Conversely, *statistical normativity* predicts intentionality ratings for positive conditions but not for negative conditions. The parameter estimates and 95% confidence intervals are presented from the hierarchical, mixed-effects model. (B) Activation in bilateral dorsal amygdala (red-yellow color map) was found to be positively associated with intentionality ratings for negative outcomes within ROIs identified from reverse inference maps of “emotion” from Neurosynth, indicated in blue. (C) This relationship was partially mediated by reports of emotion for negative consequences (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.80; 95% confidence interval = [0.07, 2.02]; Online Methods) while reports of positive emotion did not have a mediating role.  $\beta$  for separate negative and positive consequence mediation models are indicated, while the  $\Delta\beta$  indicates the change in beta value for the direct path after controlling for the indirect path. .... 102

Figure 16: Moral judgments of blame and credit serve as inputs for intentionality ascription in both *Ascription* and *Denial*. *Moral judgment* of blame served as a significant mediator of the relationship between *emotional reaction* and *intentionality* in negative conditions (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.30; 95% confidence interval = [0.18, 0.43]). *Moral judgment* of credit served as a significant mediator of the relationship between *statistical normativity* and *intentionality* in positive conditions (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.10; 95% confidence interval = [0.05, 0.13])..... 105

## Acknowledgements

I would like to acknowledge the strong support of my advisors, Professors Scott Huettel and Walter-Sinnott Armstrong. This dissertation and the training behind its formation were made possible by their mutual enthusiasm for forging bridges between neuroscience and philosophy. I am also grateful for advice, mentoring, and help through all these years from members of the Huettel lab and MADLAB. I would also like to thank my other committee members, Professors Michael Platt, Alison Adcock, and Pate Skene for helpful advice and guidance throughout graduate school. I would also like to acknowledge Professors Christopher Kontos and Dona Chikaraishi for their support for my transition into the Medical Scientist Training Program (MSTP), and I am grateful for the support of the rest of the members of the MSTP throughout all stages of my training. I would like to thank my friends, Cameron Smith and Vinal Lakhani, for fostering my interest in science through the majority of the last two decades. I would like to thank my family: Jennifer Ngo, Huong Ngo, and Long Ngo for their unconditional support through all times. Finally, I would like to thank McKenzie for being a dedicated tennis and golf partner.

# 1. Introduction

Somewhere along the borders of Ecuador, Peru and Colombia, “The Monster of the Andes” stalks his next victim. To the dismay of citizens and authorities alike, *The Monster* has evaded capture and confinement for decades. He takes great pride in his elusiveness; he gloats about being the “man of the century.” Pedro Alonso Lopez is debatably the most prolific serial killer ever documented in human history. He is singlehandedly responsible for the rape and deaths of more than 300 young girls. His body count dwarfs – by more than a factor of 10 – the number of victims associated with more notorious serial killers like England's Jack the Ripper and the U.S's Ted Bundy. His last release was in 1998 when he had been declared to be mentally sane and fit to reenter society on \$50 bail. From the early 2000s until the present, he has been responsible for additional ghastly murders, and his exact current whereabouts are unknown (Lohr, n.d.; Newton, 2000).

Zell Kravinsky also believes he is responsible for hundreds of deaths. He received two different Ph.D. degrees from the University of Pennsylvania in addition to completing the requisite coursework for a third doctorate. He then taught at the University of Pennsylvania where, according to student evaluations, he was consistently the most highly ranked professor. His good deeds were not confined by the walls of the university: he worked hard in Philadelphia to help handicapped children in the

underserved schools of the inner city. Apart from teaching and mentoring, he also proved to be a financial genius. Through a series of savvy real estate investments, he quickly amassed more than \$45 million dollars.

However, several years later, he was left with almost nothing. In the early 2000s, he gave away nearly all of this fortune to charities. Focusing mostly on charities within the public health sector, one of his donations included the largest individual contribution given to the U.S. Centers for Disease Control and Prevention (CDC). All that remained after the donation were essentials for his family: minor investments, a house, and two minivans. To him, such extreme measures were highly justified. Holding onto excess money meant withholding life-saving treatment for hundreds or thousands of ill patients.

By 2003, he still had a guilty conscience, so he hatched a plan for moral catharsis. Early one morning, he crept out of his home and drove to Albert Einstein Medical Center, unknown to his wife and children. Something that he had been secretly planning for months, he proceeded to donate one of his kidneys to a complete stranger (Parker, 2004). His guilt had been driven by a question that, to him, had ultimate moral significance: why should renal disease patients with no functional kidneys die, while he held onto his perfectly functional pair?

Stories about Lopez and Kravinsky force us to engage in a special type of cognition. It is a foundational feature our species, which we begin developing at a very young age (Hamlin, Wynn, & Bloom, 2007; Premack & Premack, 1997). A deficiency in these basic capacities result in profound disability, manifesting in mental disorders such as autism or psychopathy (Blair, 1995; Grant, Boucher, Riggs, & Grayson, 2005; James & Blair, 1996). Though it has played a large part in guiding our success as a species, the neural mechanisms of its function have only been recently examined and the picture is still unclear. In this thesis, I aim to explore an emerging model for how the brain engages in this type of cognition.

I aim to explore the nature of moral cognition.<sup>1</sup> First, I will survey the existing literature on moral cognition and highlight the unique perspectives that different researchers have taken to study it. In the second chapter, I will bring together two parts of the literature regarding unique components of morality: mentalizing and value. In the third chapter, I will take a step back and devise a methodology and framework to identify hierarchical ontologies for neural circuits relevant to cognitive neuroscience as a whole. In the fourth chapter, I will apply these methods and ontology to moral cognition, identifying the neural components of morality in a more comprehensive and

---

<sup>1</sup> All chapters in this dissertation were produced from a collaborative effort. Significant contributions to writing in Chapter 5 were made by Scott Huettel, Walter Sinnott-Armstrong, and Chris Coutlee. Across all chapters, I contributed the majority of the writing.

formal way. In the fifth chapter, I will explore the dynamics of the hierarchical model and show that morality has a top-down influence on its constituent parts. Finally in the sixth chapter, I will review the current work's implications for understanding morality with particular attention to implications in medicine.

## ***1.1 Thinking Hard about Morality for Millennia***

Two main approaches comprise the study of morality: normative and descriptive. Philosophers have studied morality from the former perspective, pondering what is right or wrong, good or bad, virtuous or vicious. In a mostly separate project, scientists have described how morality actually manifests in people and in society, leaving aside considerations of the objective nature of morality.

### **1.1.1 Normative Perspectives of Morality**

With the rise of monotheism in the west, medieval philosophers combined morality with religion in a synthesis called divine command theory. Here, theologians/philosophers such as St. Augustine (Aquinas, 1988) and St. Thomas Aquinas (Stump, 2001) proposed that morality depends solely on what God says. If God commands, "Thou shalt not steal," then that is the last word on the moral status of theft. Such logic is even taken to certain extremes: even if God commands one to murder one's own son – as Abraham was ordered to kill Isaac in Genesis – this means that it is morally right to kill.

More secular attempts attempted to summarize morality without direct reference to a deity. Immanuel Kant proposed that the myriad principles encompassed by all of moral law could be fully expressed by one foundational principle: the categorical imperative. In it, he states that one "Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction." Alternatively, he later states, "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but at the same time an end," which he finds to be an equivalent statement to the first (Kant, 2002).

Alternatively, consequentialists such as Jeremy Bentham and John Stuart Mill proposed that the principle of maximizing utility should be the foundational principle of morality (Bentham, 1907; Mill, 2002). Here, motives and character all are only good to the degree to which they ultimately lead to happiness. Such an approach becomes a major counterpoint in many empirical studies later done by psychologists and neuroscientists to Kantian morality.

Finally, a moral theory that has received increased attention within philosophy recently, but which has ancient roots in the works of Plato and Aristotle, is virtue ethics. Instead of utility or moral principles/rules, virtue ethics focuses on the nature of virtue and vice. The concepts of virtue and vice depend heavily on the "golden mean." Vices

represent extremes of deficiency or excess on a continuum of character traits; virtues occupy the desired space between these two extremes. For instance, one should try to foster courage while avoiding the alternative extremes of cowardice or rashness (Aristotle, 1959).

### **1.1.2 Psychological Perspectives of Morality**

Led by Kohlberg and Piaget, developmental psychologists were interested in the etiology of the moral sentiment. Piaget proposed that humans progress through four distinct stages of development, each of which is bound to a certain age range (Piaget, 1977). Kohlberg built on Piaget's work and proposed six stages of development from childhood through adulthood; his stages were more loosely tied to age compared to Piaget's framework. Interestingly, Kohlberg's later stages are never reached by most in the general population (Kohlberg, 1984). Turiel, a student of Kohlberg, later focused on the distinction that children draw between conventional and moral transgressions (Turiel, 1983).

In contrast to much of this developmental work, which focused on morality as a rational process, other psychologists have followed in the spirit of philosopher David Hume who famously stated that "Reason is, and ought only to be a slave of the passions" (Hume, 1978). They have identified countless emotional factors which influence moral judgment in surprising ways (Schnall, Haidt, Clore, & Jordan, 2008; Strohminger, Lewis,

& Meyer, 2011; Valdesolo & DeSteno, 2006). This sort of work has led some researchers to conclude that emotion is really the driving force behind moral judgments with rationality playing purely an epiphenomenal role (Haidt, 2001). Some of our own work on the issue, specifically regarding the nature of social conformity on moral judgments, suggests that such theories may have gone too far on their theoretical focus on emotion (Kelly, Ngo, Huettel, & Sinnott-Armstrong, 2014).

Perhaps one of the most influential recent developments in the field of moral psychology has been the consideration of whether morality is a single, unified concept. As described by Haidt in the Moral Foundations Theory (2007), a group of social and cultural psychologists have partitioned morality into six unique components, which seem to be represented consistently across all cultures: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. These researchers have pointed to the fact that much of the research on morality has focused on the first two of these components: care/harm and fairness/cheating. Though these two have been important in explaining the range of moral judgments that we all make, the others domains have been largely neglected. With the additional focus particularly on sanctity/degradation, the Moral Foundations Theory has made powerful predictions: for instance, it has been linked to attributes outside of pure moral psychology such as political orientation (Graham, Haidt, & Nosek, 2009; Haidt, 2007). Continuing with this

partitioned framework, neuroscientists have then proceeded to confirm the existence of distinct neural correlates for several of these domains - including harm, dishonesty, and disgust (Parkinson et al., 2011).

### **1.1.3 Neuroscientific Perspectives of Morality**

Taking the framework for morality established by previous behavioral work in psychology, neuroscientists have attempted to elucidate the neural mechanisms of moral cognition. Though such an enterprise has not been without criticism (Berker, 2009; Leben, 2010), much progress has been made by many researchers who have approached the problem from many different angles. This has resulted in a rough characterization of the different components of the neural processing of morality.

#### **1.1.3.1 Moral Emotion**

Consistent with the recent focus on emotion in behavioral studies on morality, researchers have mapped brain regions that seem to be associated with emotional and moral neural processing. Moll et al. (2002) found that eliciting moral emotions in subjects was associated with activation in a wide network of brain regions including the upper midbrain, orbital and medial prefrontal cortex, superior temporal sulcus, and amygdala. Though some regions are more tightly correlated with emotion than others, there seems to be consistent activation in many of the canonical "emotional" processing regions (Greene & Haidt, 2002).

The study of the role of emotion in moral judgments has been corroborated by studies in several various patient populations. Psychopaths have long been characterized as engaging in intensely antisocial behavior, exhibiting very little empathy or remorse (Hare, 2003). Such behavioral descriptions were tested using neuroimaging techniques. Researchers found that psychopaths had reduced activity in several brain regions commonly implicated in emotional processing, such as the anterior and posterior cingulate gyri and the amygdala (Kiehl et al., 2001).

In addition to thinking of emotion as a general domain, researchers focused on a particularly fascinating and influential emotion in moral judgments: disgust. Such experiments have explored how disgust may have various interpretations. For instance, garbage, sewage, and disease elicit a sense of "pathogenic disgust." This most likely evolved as a mechanism for the avoidance of substances that could lead to toxic or infectious diseases. Alternatively, disgust can be elicited by acts of sexual impropriety, such as sibling incest or bestiality. This may be because such actions may lead to a decrement in evolutionary fitness, whether through the production of maladaptive offspring or the acquisition sexually transmitted diseases. Apart from these more basic forms, disgust seems also to have a sense that is more complex. Disgust can be a common response to severe and egregious harms and societal injustice, like murder or terrorism. Despite this conceptual distinction between these "primitive" and "complex"

forms of disgust, researchers have found the two to have significant neurological overlap (Moll et al., 2005; Schaich Borg, Lieberman, & Kiehl, 2008).

Though both psychologists and neuroscientists have recently focused more on the role of emotion – as opposed to the rational approaches of earlier psychologists (e.g., Piaget, Kohlberg, and Turiel) – some research has conceived of moral judgment as a sort of competition between both perspectives. This has mainly been done within neuroscience through the revival of the centuries-old conflict between deontology and consequentialism; the particular approach involves the mapping of these two moral theories to two important neural domains often studied in cognitive neuroscience: emotion and cognitive control. The main tool for these studies was originally introduced by the philosopher, Philippa Foot, and is commonly called the “trolley problem” (Foot, 1978). A scenario based upon her formulation of the problem follows:

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Ought you to turn the trolley in order to save five people at the expense of one? (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001)

Most participants say, “yes” in this case. Judith Thomson (Thomson, 1976) later added the companion variant of the case:

As before, a trolley is hurtling down a track towards five people. You are on a bridge under which it will pass, and you can stop it by dropping a heavy weight in front of it. As it happens, there is a very fat man next to you – your only way to stop the trolley is to push him over the bridge and onto the track, killing him to save five. Should you proceed?

Here, most participants say “no.” Through a series of behavioral and neural experiments (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001), Joshua Greene and colleagues have proposed that the judgments participants make across these two scenarios boil down to the distinction between personal and impersonal judgments. The footbridge case represents a type of personal judgment, which elicits activation in several brain regions including medial frontal gyrus, posterior cingulate gyrus, and bilateral angular gyri, which Greene and colleagues claim are involved in emotion. Alternatively, Foot’s original scenario represents an impersonal judgment, which elicits activation in areas of the middle frontal gyrus and areas of the parietal lobes, which they claim to be associated with higher-order reasoning (Greene et al., 2001).<sup>2</sup>

A later study provides further analysis of the personal-type moral scenarios. While their last study found the neural correlates of exposure and engagement with personal-type scenarios, this new study employed models for personal moral judgments that take into account participants’ choices. Using this enhanced approach, the authors found that heightened activation in the dorsolateral prefrontal cortex (DLPFC) and

---

<sup>2</sup> The reverse inferences based on Greene et al.’s brain regions of activation are debatable in their accuracy and specificity for the psychological processes that they have identified. Methodologies developed in Chapter 3 and a further discussion of their implications throughout Chapter 4 and 6 will further explore the validity of these inferences. There may be reason to believe that the emotion reverse inference may better be described as reflecting the activity within the default mode network.

anterior cingulate cortex (ACC) predicted subject choices for the utilitarian judgment over the deontological judgment. They conclude that the classical philosophical debate between these two moral theories manifest in the brain as a competition between cognitive control and emotion subsystems (Greene et al., 2004).

Other research on patient populations corroborates this dual-process theory. VMPFC patients, who have certain emotional deficiencies, responded to the trolley problems in a remarkably abnormal way. Compared to normal healthy controls, VMPFC patients had higher rates of judgments consistent with consequentialism (Koenigs et al., 2007). In a different study, the same heightened consequentialist tendencies held true for another group of patients with emotional blunting: frontotemporal dementia patients (Mendez, Anderson, & Shapira, 2005).

Scenarios such as the trolley problem are best attempts to isolate the specific factors that may be relevant deontological vs. consequentialist theory. However, moral judgment involves many more factors than just these two; consequently, Greene et al.'s studies ignore much of the richness that is inherent in moral situations, judgments, and behavior. Other research has explored the greater complexity of moral judgment, studying the interplay between emotion and "higher cognition" for a variety of factors, including consequences (amount of harm), the nature of action (doing vs. allowing), and intention (intentional action vs. unintentional) action (Schaich Borg et al., 2006).

### **1.1.3.2 Moral Value**

The separate field of neuroeconomics has also elucidated aspects of the neural mechanisms moral cognition. In this relatively new field, researchers have attempted to utilize neuroscientific techniques to gain further insights into the validity and mechanisms of economic theories. However, the field also reflects a mutually beneficial relationship between the two parent fields. Providing novel insights for neuroscience, an economic perspective provides several distinct advantages particularly through the use of economic games (Camerer, 2003).

The tools and theoretical approaches of neuroeconomics address several ways in which the study of moral cognition has been limited. Previous studies on morality have used stimuli that are not easily quantifiable. Such an issue is not a problem in many neuroeconomic studies, where the inclusion of money as a core task feature allows for the use of this variable in analyses as a parametric regressor. Previous studies on moral judgment have also been limited by the degree to which task stimuli are highly abstracted from everyday experience. Such studies of morality have fundamentally been about abstract judgments and not studies of concrete decisions. An underlying reason for this is the tremendous ethical difficulty inherent in having participants make any choices of significant moral consequence (though see (Falk & Szech, 2013) for a rare attempt). Alternatively, neuroeconomic studies have made strong attempts to employ

tasks that maximize realism. Neuroeconomic studies heavily emphasize the lack of deception; and consequently, the tasks allow participants to make real-life decisions. That is, participants' decisions have real consequences: monetary gain or loss to the participant or others (e.g., charities) is at stake. The use of money in neuroeconomic studies bridges the gap between abstraction and concreteness, and within neuroeconomics, such advantages have allowed for a great insights into the brain's mechanisms for the computation of value (Levy & Glimcher, 2012) and decision-making (Platt & Huettel, 2008). The study of morality using similar tools has resulted in analogous advances.

First, there is the question of whether there is such a thing as the computation of "moral value" and what similarity it may have with similar computations for value concerning material goods. For material goods, such computations come naturally, underlying all of our behaviors in economic markets. On the other hand, the computation of value in morality seems a bit trickier, especially in the context of the great difficulty and conflict that participants have in the aforementioned trolley problems. However, Shenhav and Greene (2010) show that the neural correlates of moral value and material value significantly overlap. Expected moral value was correlated with activation in the ventral striatum; value related to moral probabilities was reflected by activity in the right anterior insula.

Because of this strong neural homology between moral and material value, it may not be surprising that the interaction between the two could be quite strong. To study this, Delgado et al. (2005) used a popular economic game called the Trust Game. In one implementation of this game, Player A gets a \$10 endowment, any proportion of which she must decide to split with Player B. Once Player B receives this money from Player A, the value of Player B's possession is increased by a factor of 3. Player B can then decide how much of this total should be returned to Player A. If Player A is maximally distrustful of Player B, then the initial move would be to give Player B none of the money and consequently leave the game with \$10 dollars. Alternatively, if Player A is maximally trusting of Player B, she would initially give Player B all ten dollars in hopes of ultimately receiving a return of \$15 dollars, assuming that Player B splits the final multiplied product of \$30 evenly. Previous studies have implicated a neural circuit that includes the caudate in the computation of value that guides behavior in this game. Consistent with these studies, the authors found this to be the case: behavior in the trust game was predicted by activity within the caudate. However, judgments of moral character interacted with signal in the caudate, reducing the association that activity in this region had with subsequent decisions (Delgado et al., 2005).

Apart from methodological contributions of neuroeconomics, there have also been connections to high-level theory. One core area of interest in neuroeconomics has

focused on the distinct neural underpinnings of model-based and model-free systems of learning (Balleine & O'Doherty, 2010; Wunderlich, Dayan, & Dolan, 2012). Model-based systems are comparatively more explicit, and for a given decision, are based on the representation of a large tree of possible consequences. Model-free systems can be more parsimonious in that they do not rely on such an extensive internal model of possible contingencies. Rather, they depend a retrospective approach where value is derived from experience of outcomes from past actions. Crockett (2013) has proposed that these two systems map onto neural computations underlying deontological and consequentialist judgments; and in many ways, there is a strong homology between Crockett's approach and Greene et al.'s dual-process theory (Greene et al., 2008, 2004; Greene, 2010).

### **1.1.3.3 Moral Mentalizing**

A third major approach in the study of morality has focused on the role of mentalizing, which is the representation of others' states of mind. As discussed further in Chapter 5, many legal traditions have made the distinction between the intent to harm (*mens rea*) and the actual harmful consequence itself (*actus reus*) (Hart, 1968). For example, the distinction between these two elements may underlie the reason for why punishments given for murder, attempted murder, and manslaughter are very different. Liane Young (Koster-Hale, Saxe, Dungan, & Young, 2013; 2007; 2009) has thoroughly

explored this important interaction. In these studies, the relevant contrasts occur between attempted harms, intentional harms, and accidental harms. Consider the following question stem: “Grace and her friend are taking a tour of a chemical plant. When Grace goes over to the coffee machine to pour some coffee, Grace’s friend asks for some sugar in hers.” For attempted harms, “Grace thinks the powder is toxic. It is sugar. Her friend is fine.” For intentional harms, “Grace thinks the powder is toxic. It is toxic. Her friend dies.” And finally for accidental harms, “Grace thinks the powder is sugar. It is toxic. Her friend dies.”

Young et al. (2007) found a special role of a specific brain region in moral judgments: the right temporoparietal junction (RTPJ). Specifically, they found the region to be active for all moral conditions, but found the highest activation of all conditions for attempted harms. The RTPJ has been implicated as a brain region specifically involved in belief attribution and social decisions (Aichhorn, Perner, Kronbichler, Staffen, & Ladurner, 2006; Carter, Bowling, Reeck, & Huettel, 2012; Saxe & Powell, 2006; Saxe & Wexler, 2005). (However, see the work of Mitchell (2008) for a counterpoint.) A later study confirmed that, for moral judgments, the RTPJ has a causal role (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). They found that RTPJ disruption led participants to judge attempted harms as more permissible. Finally, a recent study has shown that multivoxel pattern analysis (MVPA) within the RTPJ is capable of yielding

accurate moral judgment predictions (Koster-Hale et al., 2013). In the same study, such a technique was not successful in a separate group of participants with Asperger's syndrome, which has interesting implications for the study of mental disorders. This suggests that dysfunction of the mentalizing function of the RTPJ may underlie the pathophysiology of autism spectrum disorders.

Finally, a particularly clever confirmation of the model comes from a study on a special patient population. Strong assumptions of laterality underlie claims of the specificity of the RTPJ (as opposed to the LTPJ). Accordingly, the RTPJ model of mentalizing makes unique predictions about participants with disconnected cerebral hemispheres (i.e., split-brain patients). Experimenters found in a population of six patients that all made abnormal judgments, such that all were purely based on consequences rather than mentalizing (Miller et al., 2010). This, along with all the other studies, are consistent with the model that the RTPJ is specifically involved in the computations of mentalizing crucial to the formation of moral judgments.

However, all of these claims are restricted to one type of morality, which deals with the violation of moral principles warranting blame and punishment. Besides the question of what one *ought not* to do, there is also the moral question of what one *ought* to do. Some work has suggested that judgments made in this positive rather than negative sense may arise from distinct neural mechanisms (Takahashi et al., 2008). So the

question would also remain as to whether the universal claims made about RTPJ also apply to this distinct domain.

Some evidence suggests that RTPJ is indeed also associated with morality in this positive sense. Tankersley et al. (2007) found that activation elicited by the perception of agency in the RTPJ was associated with participants' reports of real-life altruism. Using a different context, a corroborating study found that gray matter volume in the RTPJ, along with its activity, was associated with altruistic behavior in an economic trust game (Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012).

The role of the RTPJ and mentalizing for moral judgments has been shown to be highly robust through countless studies. From a theoretical perspective, some have gone as far as to say that mentalizing is morality's essence (Gray, Young, & Waytz, 2012). According to this theory, emotion and value are somewhat derivative or tangential to what is uniquely concerned with moral cognition at its core. I will leave such a debate aside until chapter 4, where I will pick up this controversy using some novel meta-analytic methodologies. For now, it at least seems that this is an oversimplification, and a better understanding of moral cognition could arise from examining increasing levels of complexity. Rather than viewing each of the components of morality described above in isolation, intriguing insights can be found by looking at their interactions.

#### **1.1.3.4 Moral Component Interactions**

For the most part, the way in which the neural components of morality have separately been described in this dissertation point reflects the way in which work in the field has been conducted. Different researchers have approached the problem from different angles, illuminating the nature of moral judgment in unique ways. Less work has focused on the integration between these domains; but there have been a few exceptions.

At the intersection of mentalizing and emotion, some researchers have studied a different type of mentalizing. In the studies described above, the relevant type of mentalizing was the attribution of beliefs to other agents. Often, this type of mentalizing is more relevant to the moral judgments of potential moral villains, which lead to computations of moral blame and punishment. On the other hand, another group of researchers have been interested in the mentalization of emotional and sensory mental states. In relation to moral judgments, these types of considerations would be more relevant to judgments and behaviors regarding the victims of moral harms.

Operationally defined as empathy, Singer et al. (2004) studied the neural mechanisms of feelings directed towards the victims of harm. Previous theories had proposed that empathy relies on the simulation of others' emotional and sensory via the engagement of the same neural circuits that would be involved in the processing of

one's own emotional and sensory states (Preston & de Waal, 2002). Contrary to this theory, Singer et al. found that there was a significant dissociation between the neural mechanisms of empathy and one's personal experiences. Neural activity associated with knowing of a loved one's pain activated bilateral anterior insula, rostral ACC, brainstem and cerebellum, while the direct experience of a pain stimulus activated posterior insula, sensorimotor cortex, and caudal ACC. Based on these results, the authors conclude that empathy activates only a subset of the total set of regions often implicated in the experience of pain, and these empathy-related regions are specific to higher-order representations of the subjective aspects of pain processing.

At another relevant intersection, Hsu et al. (2008) studied the interaction between emotion and value. To do so, their task employed a classic tension within moral/political philosophy between equity and efficiency. The core problem in this conflict is the trade-off that people and policy-makers must make in ensuring that resources among a population are distributed equally; and often, this comes at the cost of efficiency. Hsu et al. hypothesized that the opposing forces underlying this conflict come from emotion and value neural subsystems. Equity would be encoded by emotion, and deviances from optimal fairness would elicit negative emotions such as anger and outrage. Alternatively, efficiency would be encoded by value processing regions. Their

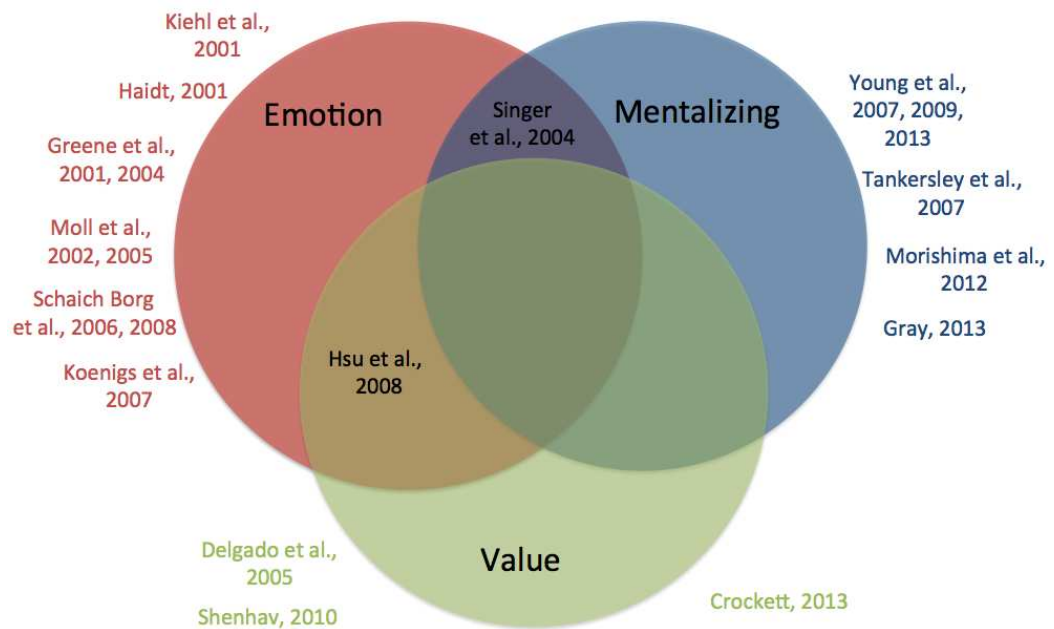
hypotheses were confirmed by their results. They found that considerations of equity were encoded in the insula while efficiency was encoded in the putamen.

In broad terms, the value-emotion interaction studied by Hsu et al. (2008) represents a different type of inter-component relationship than that studied by Singer et al. (2004). For Hsu et al., the relationship between emotion and value are very much a direct conflict: final behavior is determined by the outcomes of subsystem competition. This competition is similar to that proposed by Greene and colleagues (Greene et al., 2008, 2004, 2001; Greene, 2010), which focuses on the conflict between emotion and cognitive control. On the other hand, the mentalizing-emotion connection represented by empathy is a case in which one component (mentalizing) employs the computation of another component (emotion).

The difference in type of interactions possible between these components highlights the great diversity of interactions between components that remain to be explored. In other words, the interactions between the three components we have briefly explored in this chapter – emotion, mentalizing, and value – are not fully accounted for in terms of simple permutation. In fact, the interaction between any of these components (not to mention among all three) can, in themselves, be widely diverse.

However, significant gaps still remain in the intersection of these components (Figure 1), and a fuller account of how the brain handles these computations is essential

in understanding more complex behaviors. In the next chapter, I will attempt to fill one gap between value and mentalizing using a complicated decision making task that allows for isolation of both mentalizing and value components. Specifically, I will explore how these two brain systems seem to represent two different aspects of the same task. And to do so, I will explore the neural mechanisms of a decision-making process that has rapidly gained more attention the past few years with the implementation of healthcare reform in the United States.



**Figure 1: Survey of major work in the literature pertaining to moral judgments and decisions across the three domains of emotion, mentalizing, and value.**

## **2. Mentalizing and Value in the Disclosure of Conflicts of Interest**

### **2.1 Introduction**

In the United States, the Physician Payments Sunshine Act requires the collection and tracking of all financial transactions between physicians and industry. In 2010, this provision was signed into law by Barack Obama as part of the larger Patient Protection and Affordable Care Act. Though many of the details of the act have not yet been implemented, the intended goal of these changes is to protect patients from conflicts of interest. Such conflicts can arise from financial relationships between their healthcare providers and outside interests such as pharmaceutical companies. However, there is great complexity in the interplay among physician, patient, and outside interest, and this complexity could complicate such policies' consequences. Will mandatory disclosure have a positive impact on the state of care for patients, or could there be adverse consequences contrary to the lawmakers' intentions?

Previous work suggests that disclosure policies could have adverse effects (Cain, Loewenstein, & Moore, 2005, 2011; Loewenstein, Sah, & Cain, 2012). They have shown that disclosure policies may lead advisors to give even more biased advice compared to otherwise similar contexts without such a policy (Cain et al., 2005). There seem to be two main explanations. The first, *moral licensing*, is based on previous research outside the domain of advice that demonstrates that engaging in moral behavior can cause

participants on subsequent tasks to feel licensed to act immorally (Monin & Miller, 2001). In the advice-giving domain, the act of disclosure may be represented as an instance of this moral behavior, which then gives the advisor license to give advice that is more biased. The second reason is *strategic exaggeration*: advisors consider the effects that the disclosure of their conflict of interest will have on the decision-making process of the advisee; the advisor then compensates for possible advisee discounting of advice by exaggerating the initial degree of bias.

However, the medical context represents a unique domain for advice giving. More so than many other settings for advice, such as in the financial domain, medical advice may have direct consequences on health and illness. The professional duties of a physician may represent a special case. Physicians traditionally engage in a highly moralized ceremony before engaging in the practice of medicine where they vow to “never do harm” in the Hippocratic Oath. What sort of influence do disclosure policies have in this highly moralized context?

The purpose of this chapter is two fold: (1) Understand what influence disclosure policy has in the medical context. (2) Understand the neural mechanisms for this influence. More specifically, how do the components of the neural processing of morality – emotion, mentalizing, and value – interact to generate behaviors in this case? Elucidating the composition of this decision-making process could ultimately lead to

hypotheses and predictions about what sorts of factors could be most helpful in facilitating the physician-patient relationship in the context of conflict of interest.

## **2.2 Behavioral Methods**

### **2.2.1 Participants**

Thirty-three adults (mean age 24 years; range 18-35 years; 20 females) participated in the study. Three participants were excluded for technical issues related to data acquisition at the time of scanning. All participants provided written informed consent as part of a protocol approved by the Institutional Review Board of Duke University Medical Center.

### **2.2.2 fMRI Task**

The participants engaged in three runs of 32 trials each for a total of 96 trials per subject of a clinical advice-giving task (Figure 2). They were instructed to assume the role of physicians in a clinical setting, and their jobs were to give advice to patients. The physicians were to recommend one of two treatments. Each treatment had different monetary consequences for both parties (Figure 2).

The consequences varied across three distinct types of trials. In *conflict* trials, the choice of treatment A would result in the recommendation of better outcome for the physician but the worse outcome for the patient. The opposite was true for treatment B, where patient would have received the recommendation for the better option while

physician would receive the worse outcome. 52 out of 96 total trials were this trial type. In "non-conflict" trials, treatment B was always the better option for both physician and patient. 26 out of 96 were *non-conflict* trials. Finally, the remaining 18 of 96 trials were *ambiguous* trials, where the expected utility for both parties for both treatments were matched, and the treatments only varied in the range of possible outcomes.

Disclosure policy was expressed in two different stages during each trial. During the first screen of each trial, a symbol was displayed which corresponded to one of two different *clinics*. The *Open Atlantic* clinic mandated that physicians disclose a conflict of interest to the patient later in the trial, while the "Shielded Pacific" clinic meant that there was no disclosure policy at any point in the trial. Later, during the decision screen, a trial in the *Open Atlantic* clinic included a statement of disclosure, which was displayed to the physician, and which would subsequently be displayed to a separate participant filling the role of patient in a different session. A trial in the *Shielded Pacific* clinic omitted such a statement. Half of all trials of each consequence type were assigned to each type of disclosure policy (Figure 2).

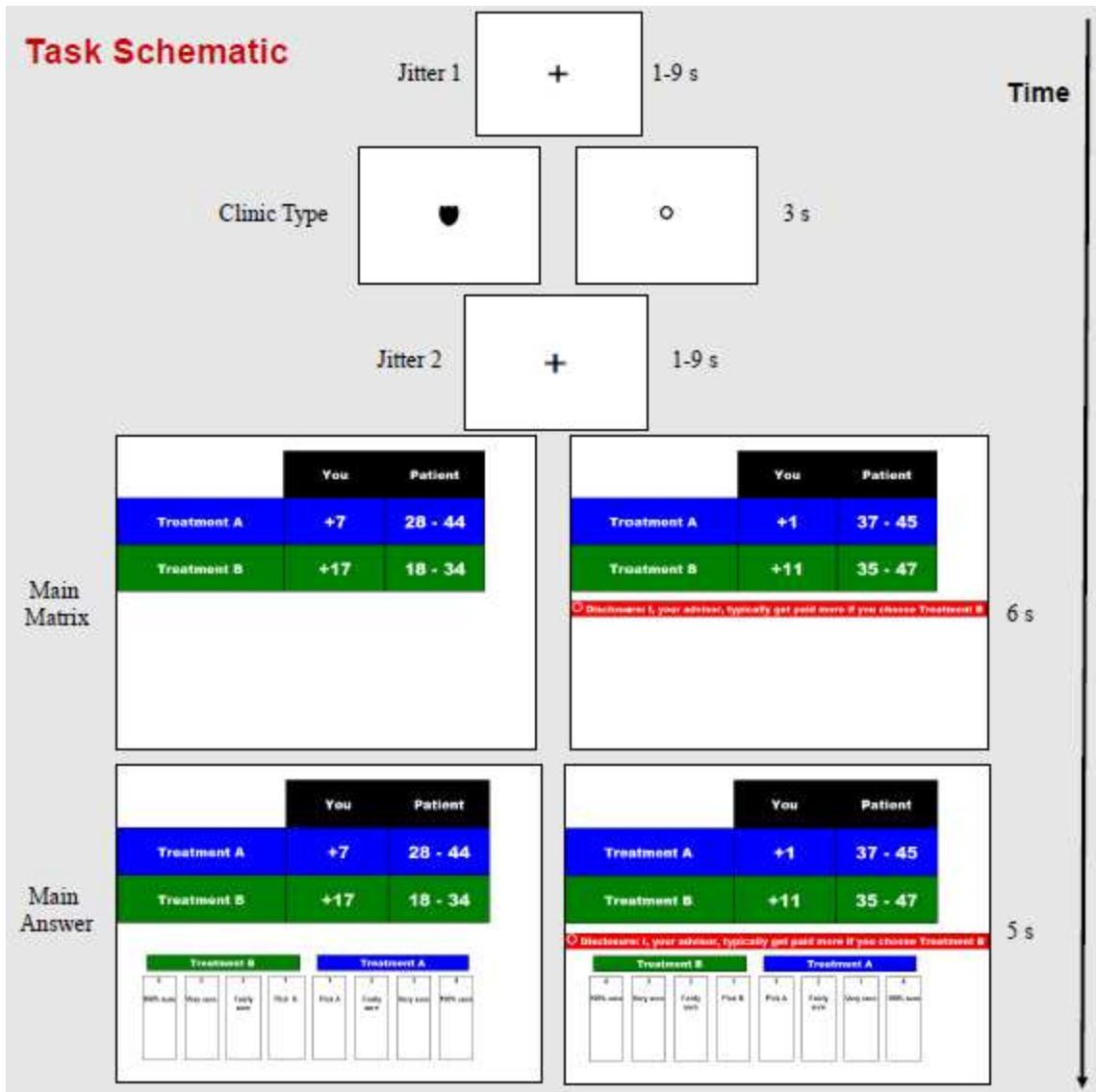


Figure 2: Task structure. After the experiment, one trial was chosen at random, and the physician's advice will be given to different participant. Both physician and patient are paid according to the option chosen by the patient in virtue of the received advice in the form of a gift card.

Subjects also responded to two separate post-scan, individual difference measures: the Cognitive Reflection Task (CRT) and the Machiavellian Scale (Mach-IV). The CRT is a measure of the disposition to spend time and resources to reflect on a problem and resist settling on first intuitions (Frederick, 2005). The Mach-IV, which is part of the dark triad, measures the degree to which participants report a manipulative attitude that is consistent with the writings of Niccolo Machiavelli, as summed up by his phrase, "The Ends Justify the Means" (Christie & Geis, 1970).

### **2.2.3 Analysis**

After data collection, it was discovered that the orientation of the responses were not randomized for left-right directionality in one of three runs. Because of considerations for scanning, this run was eliminated for each subject. The final analyses included two runs for each subject. Further, additional regressors were included for orientation of scale in neuroimaging analyses (see below). Hierarchical mixed-effects analysis (interactions factorial to 2nd degree) using subject as a random effect and conflict type and disclosure policy as fixed effects were performed in JMP 10 using the residual maximum likelihood (REML) method.

## **2.3 Imaging Methods**

### **2.3.1 Acquisition and Preprocessing**

Functional data was acquired using a 3T GE scanner with an 8-channel receiver using a spiral-in sensitivity encoding sequence. Three runs of 444 time points were acquired with TR=1.58, TE=30ms, voxel size=3.8mm x 3.8mm x 3.8mm, field of view=243mm, and flip angle=70°. Due to a coding error in the task (described above), only two runs from each subject were included in the analyses. Brain tissue was isolated using the brain extraction tool (Smith, 2002). The first eight volumes of each analyzed run were discarded to account for magnetic stabilization. Differences in slice acquisition times were corrected using Fourier-space phase shifting. Spatial smoothing was performed with a Gaussian kernel with a full width at half maximum of 6mm. Grand mean scaling was performed across datasets from each run of each subject. A high-pass temporal filter was applied with a Gaussian-weighted least-squares straight line fitting with  $\delta = 100$  s. Functional images were registered to subjects' high-resolution structural images with FLIRT, and subsequently, to MNI standard space with FNIRT (Jenkinson, Bannister, Brady, & Smith, 2002). Head motion was corrected by realigning the time series to the middle volume using FLIRT (Jenkinson et al., 2002).

### 2.3.2 fMRI Analysis

All fMRI analyses were performed with FEAT (fMRI Expert Analysis Tool) Version 5.98, which is part of FSL (FMRIB's Software Library). Time-series local autocorrelation correction was carried out with FILM (Woolrich, Ripley, Brady, & Smith, 2001). For the clinic epoch, the first-level (within-run) analysis had two regressors (disclosure and non-disclosure). For the main matrix epoch, there were three categorical regressors (conflict, non-conflict, and ambiguous). The conflict and non-conflict categorical regressors each had four parametric regressors (disclosure vs. non-disclosure, key orientation, difference in payoff between treatments for patient, and difference in payoff between treatments for advisor) occurring simultaneously during the main matrix epoch for an overall total of 13 regressors. The key orientation was included in the model to regress out any potential confounds due to unbalanced key orientation. Second-level analyses (across-runs, within-subjects) used a fixed-effects model, and third-level analyses (across-subjects) used a mixed-effect model (FLAME 1). All images presented are z-statistic images were thresholded using clusters obtained with  $z > 2.3$  and a corrected cluster-significance threshold of  $p < 0.05$  (Worsley et al., 2002).

### 2.3.3 Neurosynth Decoding

For the reverse inference terms provided in Table 3, relevant clusters of interest were submitted to the *decode* utility included in the Neurosynth Core Tools Package (Yarkoni, 2013). The utility iteratively runs a correlation between an SPM of interest and the whole-brain reverse inference maps of all 525 terms included in the database. These submitted clusters included the DMPFC cluster from the conjunction analysis from Figure 1 and the ventral striatum cluster from Figure 2.

### 2.4 Behavioral Results

According to a hierarchical mixed-effects analysis, there was a main effect of conflict vs. non-conflict ( $\beta=-0.85$ ,  $SE=0.19$ ,  $t=-4.45$ ,  $p < 0.0001$ ) such that participants chose option A more often when there was a conflict of interest ( $M=5.16$ ) compared to no conflict of interest ( $M=6.88$ ). This means that for non-conflict trials, participants choose the rational option that maximized payoff to both self and patient. Participants modified their behavior based on whether there is a conflict of interest as indicated by increased recommendations in favor of patients (treatment A). There was also a main effect of disclosure ( $\beta=-0.23$ ,  $SE=0.09$ ,  $t=-2.61$ ,  $p = 0.01$ ), such that participants chose option A more often when they were required to disclose their conflict of interest.

Importantly, there was a significant interaction between conflict and disclosure ( $\beta=-0.16$ ,  $SE=0.04$ ,  $t = -4.44$ ,  $p < 0.0001$ ). In a separate model to interrogate this interaction,

we included only conflict trials in a hierarchical, mixed effects analysis with participant as a random effect. This restricted model demonstrated that there was a significant effect of disclosure ( $\beta=-0.39$ ,  $SE=0.11$ ,  $t=-3.44$ ,  $p < 0.001$ ), such that participants provided advice that was better for the patient under disclosure vs. non-disclosure policy for conflict trials.

## **2.5 fMRI Results**

All of the following presented data identifies BOLD response in brain regions during the main-matrix epoch according to random-effects, whole-brain regression analyses. Additionally, the model has regressed out any potential confounds due to unbalanced key-orientation presentation.

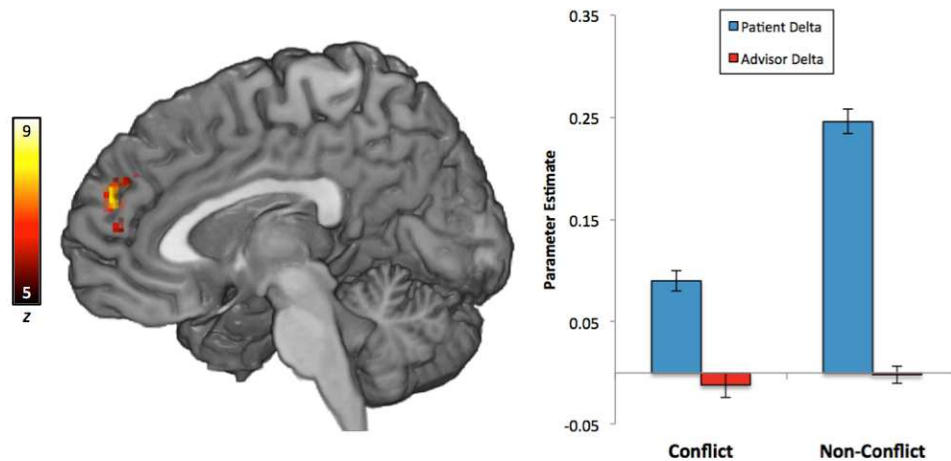
### **2.5.1 DMPFC tracks the patient outcomes**

For conflict trials, the BOLD response in a region in the dorsomedial prefrontal cortex (DMPFC) tracked the difference in available outcomes between the two treatment options for the patient, where the larger payouts to the patient for treatment A relative to treatment B were associated with greater BOLD response. Similarly on non-conflict trials, a largely overlapping region of the DMPFC also tracked the difference in patient payouts. The intersection of significant voxels across these two analyses is presented in Figure 3 (max voxel, MNI: (4, 54, 26),  $z = 8.37$ ). No brain regions significantly tracked the difference in expected payouts to self (physician) in conflict or non-conflict trials (Figure

3). Using the *decode* utility from the Neurosynth Core Tools Package (Yarkoni, 2013), we identified the features from the Neurosynth database that had the highest correlation with the cluster of activation from this DMPFC cluster of activation, which included *moral, mentalizing, intentions, and tom* (Table 3).

**Table 1: Clusters of activation for the intersection of voxels that significantly tracked the level of conflict of interest for both conflict and non-conflict trials. All indicated clusters passed a voxel significance threshold of  $z > 2.3$  and were whole brain corrected at  $p < 0.05$ .**

Region	Voxels	x	y	z	Z-max
Dorsomedial Prefrontal Cortex	84	4	54	26	8.37
Medial Prefrontal Cortex	15	6	52	10	6.29
Dorsomedial Prefrontal Cortex	4	2	44	32	6.01
Dorsomedial Prefrontal Cortex	3	-18	58	12	5.94



**Figure 3: Dorsomedial prefrontal cortex tracks the difference in possible patient outcomes. (A) The intersection between significantly active voxels found to significantly track the magnitude of difference between patient payments for conflict and non-conflict trials yielded a cluster of activation in the dorsomedial prefrontal cortex (whole-brain corrected at  $p < 0.05$ ; max voxel, MNI: (4, 54, 26),  $z = 8.37$ ). Heightened activation in this region was associated with larger differences between outcomes in Treatment A and B for the patient. (B) Parameter estimates drawn from the region shown in (A) are significantly positive for differences in outcomes for the patient. However, those drawn from the parameter estimates of outcomes for self were not significantly different from zero. ( $n = 30$ )**

## 2.5.2 Striatal activation predicts individual differences in responsiveness to disclosure policy.

In measuring the influence of a disclosure policy on decision-making, the behavioral metric of interest (*disclosure payoff difference*) was the difference in total payoff of chosen treatments for self vs. patient across the two disclosure conditions. Activity in the striatum, along with clusters in occipital cortex and thalamus, significantly predicted individual differences in response to disclosure policy for conflict trials (Figure 4).

Subjects exhibiting greater activation in these regions for the disclosure > non-disclosure contrast tended to change their behavior in favor of recommending better treatments for patients in disclosure trials compared to non-disclosure trials. Notably, the same relationship did not hold for non-conflict trials for this region in ventral striatum (Figure 2). Using the same *decode* utility as above, we identified Neurosynth features correlated with the cluster of activation in the ventral striatum, which included *reward*, *outcome*, and *monetary* (Table 3).

**Table 2: Maxima for clusters whose activity significantly tracked the level of patient-biased responses to disclosure policy. All indicated clusters passed a voxel significance threshold of  $z > 2.3$  and were whole-brain corrected at  $p < 0.05$ .**

Region	Voxels	x	y	z	Z-max
Inferior Occipital Cortex	3366	2	-64	4	3.7
Superior Occipital Cortex	3152	-20	-58	44	3.57
Thalamus	1125	-4	-14	10	3.67
Ventral Striatum	705	-6	20	-8	3.36

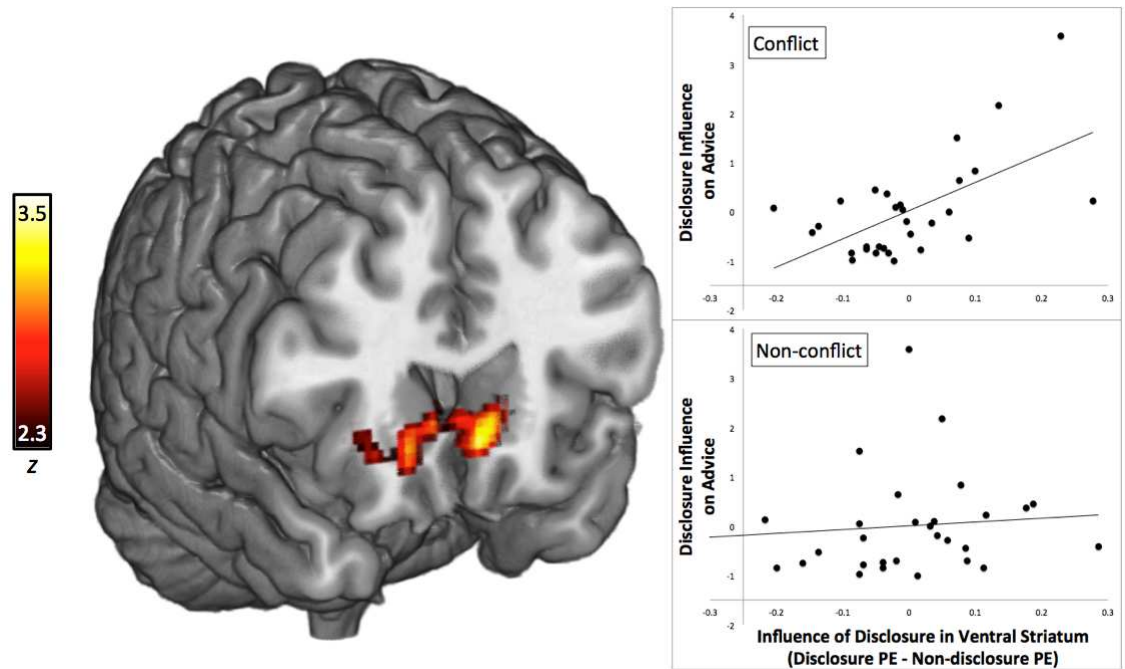


Figure 4: Striatal activity is associated with response to disclosure policy for conflicts of interest. (A) Individual differences in behavioral response to disclosure policy for conflict trials is positively correlated with activity in the striatum (whole-brain corrected at  $p < 0.05$ ; striatal peak, MNI:(-6, 20, 08),  $z = 3.36$ ). (B) The same comparison for non-conflict trials was not significant ( $R = 0.095$ ;  $p = 0.61$ ). The scatter plot for conflict trials recapitulates the data in (A) and is provided for illustrative purposes. The “Disclosure Influence on Advice” metric is a difference of differences measure. First, the difference in accumulated payoffs of given advice for trials between patient and self are calculated separately for disclosure and non-disclosure trials. The difference is then taken between these two conditions. The “Influence of Disclosure in Ventral Striatum” is the difference in parameter estimates across disclosure and non-disclosure trials drawn from the ventral striatum ROI shown in (A). ( $n = 30$ )

**Table 3: Top ten Neurosynth terms associated with clusters of interest from Tables S1 and S2. This includes all of the clusters in Table 1 for DMPFC and the ventral striatum cluster from Table 2.**

Term	Correlation
<b>DMPFC Tracking Patient Payments</b>	
moral	0.071
story	0.069
scenarios	0.066
mentalizing	0.065
social	0.060
default	0.060
mental	0.059
intentions	0.058
person	0.056
tom	0.056
<b>Ventral Striatum Tracking Disclosure Receptiveness</b>	
reward	0.143
outcome	0.139
rewards	0.132
monetary	0.132
incentive	0.126
anticipation	0.121
outcomes	0.118
impulsivity	0.100
choices	0.098
money	0.095

## **2.6 Discussion**

Contrary to previous literature on disclosure of conflict of interest, we have found that participants modify their behavior in favor of patient-centered outcomes in virtue of a disclosure policy for conflicts of interest. Further, we found that two major aspects of our task depend two distinct neural subsystems. The DMPFC, which has been implicated in mentalizing, tracks the magnitude of the conflict of interest. Alternatively, the striatum tracks the degree to which participants modify their behaviors in favor of the patient due to a disclosure policy.

The divergence of our participants' behavior may be due to the framing of our task, particularly with regard to its moral nature. Previous studies have employed more strategic and non-moral contexts in the exploration of the possibly perverse effects of disclosure. For example, Cain et al. (2005) use a task where an estimator must make a guess about the amount of money contained in a jar of coins. The advisor was another participant who had a better view of the jar of coins; and with this informational advantage, the advisor made recommendations to the estimator. The advisor earned more money if she was able to convince the estimator to overestimate the value of coins, while the estimator earned money in proportion to the accuracy of the guess. Compared to our study, the framing of this previous task was highly strategic and game-like. Participants may not have had any real motivation towards altruism in the task, and in

fact, they may have had competitive motivations to win the “game” at cost to the advisee.

Several features of our task reinforce heavy associations with a real healthcare context. Subjects are explicitly told to assume the role of physicians who are advising patients on possible treatments options. The disclosure policies are framed in terms of different rules corresponding to different medical clinics. Upon debriefing, many of participants were either trainees or had the intent to work within the healthcare field. Perhaps most importantly of all, participants completed our study in the imaging facilities of a large hospital. The radiology suite was located centrally in the hospital, meaning participants had substantial exposure to physicians and patients right before the study.

Because of these factors, we hypothesize that the uniquely moral nature of the professional duties of a physician (as opposed to, say, financial advisor) influenced our participants to behave in a way that was categorically different from previous studies. This means that the theories for mechanism of perverse influence of disclosure in conflicts of interest did not apply and suggests that the effects of moral licensing or strategic exaggeration are largely diminished within this medical context.

Neurally, our results demonstrate that the sensitivity of the DMPFC is higher for differences in patient outcomes than they are to advisor (self) outcomes. This is consistent with previous meta-analytic work showing an increased relevance of other-related neural processing as one moves within the prefrontal cortex from ventral to dorsal aspects (Denny, Kober, Wager, & Ochsner, 2012). On the other hand, our findings are inconsistent with previous theories on the perverse effects of disclosure. Theories such as moral licensing and strategic exaggeration would make the opposite prediction that more neural resources would be dedicated to outcomes for self rather than patients. Contrary to this, our results suggest that the moral context of our task has shifted the focus of processing, consequently influencing advisors' behaviors in a way favorable to patients.

Concerning disclosure policy, the striatal result suggests several possible interpretations. It could be that some feature of the scenario, perhaps the prior disclosure screen, may itself be rewarding. Other research has demonstrated that the act of disclosing information in a general sense activates regions associated with the brain's reward circuitry, including the ventral striatum and the ventral tegmental area (Tamir & Mitchell, 2012). Disclosing a potential moral conflict could also be rewarding for our participants, and the degree to which this reward is reflected in the ventral striatum could lead to more patient-biased advice. Alternatively, the reward-related activity

within the striatum could be the consequence of the patient-biased behavior itself. Here, this activity would be analogous to the “warm-glow” often described in the context of altruism (Harbaugh, Mayr, & Burghart, 2007).

Unlike the inter-component relations described in Chapter 1, the relation between mentalizing and value in this task seem to represent a third of class of interaction. Where the work of Hsu et al. (2008) represented a competition between value and emotion subsystems and that of Singer et al. (2004) represented the use of the emotional subsystem for a computation relevant to the mentalizing subsystem, our results from this study suggest that mentalizing seems to track the overall context of a situation (particularly in relation to the consequences relevant to the patient), while the value system is more intimately tied to the influence that disclosure has on the final behavior.

However, all of these conclusions depend on several reverse inferences, which may be particularly difficult in the case of tying the DMPFC to mentalizing. The use of data from Neurosynth helps to provide more precision in the strength of reverse inference through a Bayesian algorithm (Yarkoni, 2011). In doing so, Neurosynth makes use of a large and comprehensive set of imaging studies to calculate metrics of specificity, which takes the traditional yet informal model of reverse inference to a higher level of precision. Further, we employed a tool with the Neurosynth Core Tools

Package (Yarkoni, 2013) called the *decoder*, which aids in making reverse inferences for a group of voxels rather than at single voxels as implemented on the Neurosynth website. Even in using these techniques, there is still some weakness to the reverse inference that we make in Table 3. Though most of the top brain maps belong to terms that have something to do with mentalizing, we cannot rule out with great confidence that the activation in our study does not represent default mode network, which also is represented in Table 3. Instead of mentalizing, this would suggest an alternative reverse inference of resting state.

In the next chapter, we will develop novel tools – that among other uses – will aid in increasing the precision of the reverse inference that we have presented in the current chapter. The implications of these findings will be presented in greater detail in Chapter 6. Further, we have been working with an informal model of the neural components of morality as determined by a survey of the literature. Perhaps researchers could be using invalid reverse inferences, while other relevant components of morality could be understudied because of biases in the literature. In the next chapter, we also aim to implement a more data-centered survey of the literature in the attempt to provide a more formal conceptualization of what the neural components of morality really are.

## **3. The Neural Elements and Compounds of Cognitive Neuroscience**

### ***3.1 Introduction***

In 2011, a computer system named Watson handed defeat to two of the strongest human players in the history of the game show, Jeopardy. In the same year, the U.S. state of Nevada passed a law allowing for the operation of autonomous, driverless cars. As technology advances, the capabilities of computers will continue to advance, infringing on tasks that have always previously been viewed as uniquely human. Considering that many tasks are now being automated, what unique qualities does the human brain have?

We propose that one answer lies in the human's brain capacity for versatility. Even though it is dwarfed in size and mass by the most cutting-edge computer systems, the brain is able to do an incredible number of tasks with only 3 pounds of matter. So even though the humans lost to Watson in Jeopardy, they were able to easily drive home, while Watson will forever be confined to its server room. Similarly, the autonomous cars in Nevada have very little chance of ever performing well on Jeopardy. Game shows and driving surely are not the only cases: humans can do an almost infinite number of tasks ranging from solving math problems, experiencing emotions, making moral judgments, and committing things to memory.

Despite the limited number of neurons and space inside the skull, we propose that the brain can do many things because of combinatorics. An apt analogy is the periodic table of elements from chemistry. Chemists have identified 114 unique elements in nature, which represent the 114 different types of atoms that exist. These elements can combine in many different ways to form molecules, and the number of molecules possible, composed of combinations of elements, is infinite. From just a few elements, combinatorics yields the vast diversity of matter contained in the universe.

Similarly, we hypothesize that patterns of neural activation are also composed of basic neural elements. And to achieve the great diversity of human sensation, thought, and behavior, these basic neural elements combine in different ways to form a plethora of emergent *neural compounds*. Accordingly, there are two major aims of this chapter. The first will be to identify the basic *neural elements* underlying human cognition. Once this is established, the second aim will be to analyze the elemental compositions of various *neural compounds*. The source of our data for this project will come from Neurosynth, which is the largest meta-analytic database of human fMRI studies available at present.

## **3.2 Materials and Methods**

### **3.2.1 Neurosynth**

We obtained our raw data from Neurosynth, which is a publicly available database of automated meta-analyses of over 5800 fMRI studies on a diverse range of

topics within cognitive neuroscience (Yarkoni et al., 2011). The data was formed by analyzing the frequency of various terms in the body of these fMRI studies. For instance, articles that use a certain term, e.g., pain, above a rate of 1/1000 words are tagged as articles pertaining to pain. Additionally, the coordinates from the neural activation tables of these pain-tagged studies were extracted. The process was repeated across hundred of terms of interest within cognitive neuroscience, and various Bayesian algorithms were applied to produce meta-analytic brain maps.

Of particular note are the reverse inference maps that are available through the website or through the Neurosynth Core Tools Package, which allows for customized production of these reverse inference maps. The reverse inference maps are formed from a Bayesian algorithm that calculates the posterior probability that some term is used in an article given that the article reports activation at a certain coordinate. This allows for a measure of specificity. For example, if a certain brain region were associated with activation with every term in the database, the reverse inference map would account for this and yield a low value for any term at this location.

At the time of this writing, there are a total of 525 terms available in the Neurosynth database. However, many of these terms are semantically equivalent to one another. A thesaurus was accordingly constructed and utilized to merge similar terms within the database. A very conservative threshold was used for this process. Merged

terms fell into one of two major categories: inflected variants (e.g., “action” and “actions”) and English variants in spelling (e.g., “color” and “colour”). To merge terms, we generated meta-analytic maps representing the union between two or more terms. For example, the merging of “action” and “actions” results in a meta-analysis that identifies studies that use either “action” and/or “actions” at a rate higher than 1 per 1000 words. For the reporting of data throughout the paper, preference was given for inflected variants with fewer suffixes (usually the lemma) and American English variants. We were left with 414 remaining terms after merging. The Neurosynth Core Tools Package (Yarkoni, 2013) was used to generate meta-analytic, unthresholded, reverse inference maps for each of these terms.

### **3.2.2 ICA Decomposition**

We then performed Independent Components Analysis (ICA) on the entirety of the remaining database. The 414 reverse inference images were concatenated to produce a four-dimensional dataset such that the first three dimensions represented space and the fourth represented a unique Neurosynth term. Analysis was carried out using Probabilistic Independent Component Analysis (Beckmann & Smith, 2004) as implemented in FSL's Multivariate Exploratory Linear Decomposition into Independent Components Version 3.10 (FMRIB, 2014). The following pre-processing was applied to the input data: masking of non-brain voxels, voxel-wise de-meaning of the data, and

normalization of the voxel-wise variance. Pre-processed data were whitened and projected into a 65-dimensional subspace. Principal Component Analysis (as in Smith et al., 2009) was then performed on the dataset. The whitened observations were decomposed into sets of vectors, which describe signal variation across the temporal domain (Neurosynth term) and across the spatial domain (maps) by optimizing for non-Gaussian spatial source distributions using a fixed-point iteration technique (Hyvärinen, 1999). Estimated Component maps were divided by the standard deviation of the residual noise and thresholded by fitting a mixture model to the histogram of intensity values (Beckmann & Smith, 2004).

After dimensional decomposition using ICA, we found there to be two distinct types of components: task-related and participant-related. For theoretical purposes discussed in more detail below, we have excluded the participant-related terms from further analyses. However, we included them in our original ICA analyses to examine whether such terms could be separated from task-related terms using our data-driven approach.

### **3.2.3 Descriptive Thresholding**

The ICA decomposition yields pairs of spatial map and “time course.” The spatial map denotes the three dimensional structure of a given component, while the “time course” represents the relevance of that spatial map across the fourth dimension,

which in our case, is the Neurosynth term. So all 414 Neurosynth terms have a measure of relevance for each of the components from the ICA analysis.

With “time course”-related measures of relevance, we provide a succinct description of each of the ICA components using Neurosynth terms. We ordered all of the Neurosynth terms from greatest to least relevance for each component. In this descending list, we then calculated the differences between neighboring components, and a descriptive threshold was set at the maximal difference point. In other words, the threshold was set at the maximal value of the first derivative of the  $f(x)$ , where  $x$  is Neurosynth term within the descending relevance list and  $f$  is the mapping of each Neurosynth term to its relevance to each ICA component, as indicated by the “time course” matrix from the ICA decomposition. Only Neurosynth terms that occur above the maximal threshold on the descending list of terms is presented for each of the accompanying components for subsequent visualization.

### **3.2.4 Clustering**

Also for visualization purposes, we clustered the ICA components based upon their similarities in ICA decomposition data. First, we concatenated the “time courses” of all of the 50 task components after removing participant-related, artifactual components, and negative components, as described in more detail below. We then calculated a Euclidean distance matrix by representing each of the components as a

unique observation in 414-dimensional space; each value corresponds to relevance values from the concatenated “time courses.” This distance matrix was then used for hierarchical clustering using a Nearest Point Algorithm as implemented by the Python Scipy Cluster Hierarchy Package (“Scipy,” 2013). The resulting clusters (flattened at experimenter-determined levels) are represented within Figure 5, while ICA components that were not reliably clustered with any other components were arranged in the same figure at the researchers’ discretion for visualization.

### **3.3 Results**

#### **3.3.1 Component Categorization**

The automatic estimation algorithm as implemented by MELODIC (FMRIB, 2014) yielded an optimal decomposition into 65 components. These components were further classified into groups, depending on various criteria. First, we found a group of seven components with spatial patterns corresponding roughly to the gray-white matter junction throughout the brain. These artifactual components were set aside in further analyses, which is consistent in method and proportionality with previous work (Smith et al., 2009).

Second, there were a group of three components that reflected participant-relevant brain differences. In contrast to task-related terms, the Neurosynth terms most relevant to these components relate to the participants of the contributory studies and

include *sex, women, autism, adults, and elderly*. Importantly, such terms were only strongly associated with the three participant-related components and not to any other task-related components. The participant-related components were also omitted from further analysis and visualizations.

Of the remaining components, there were four networks that did not uniquely load onto any number of terms positively. Instead, they were relevant to the components in a negative way. For instance, component 36 was best characterized as being everything except for *visual*, 59 was not *time, image, and response*, 61 was not *decision*, and 63 was not *engaged*. The significance of these networks remains open for further interpretation, but for the purposes of visualization, we have also excluded these networks.

### **3.3.2 The “Periodic Table of Neural Elements”**

After the elimination of *artifactual, participant-related, and negative components*, we were left with 50 components that we will call the *neural elements*. They are presented in Figure 5. In general, the figure arranges the elements from left to right in terms of relevance to sensory and motor processing. Within the figure, some elements have more descriptive Neurosynth terms than others as a result of our method of descriptive thresholding. The arrangement of the elements was partially determined by the hierarchical clustering procedure. This procedure tended to cluster elements that had a

low numeric designation (i.e., components from the ICA that explained more variance) to a higher degree than those with high nominal numbers. The ICA decomposition resulted in the separation of certain components from one another, such as element 13 (*tom, mental, story, social, mentalizing*) and element 26 (*default, rest, restingstate*). But at the same time, hierarchical clustering indicated that such terms, though distinct, were still related.

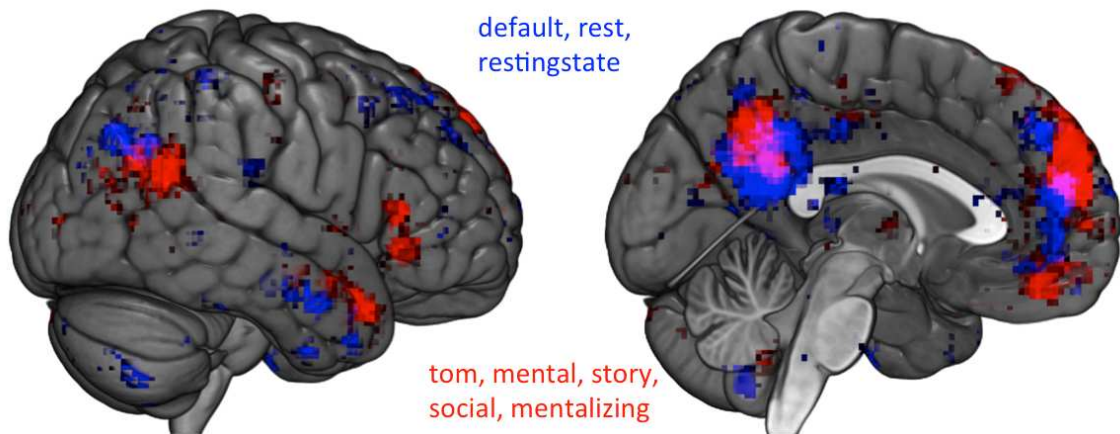
### **3.3.3 Brain Region Parcellation**

The spatial maps of the neural elements were found to overlap, though the ICA decomposition minimized its degree. The comparison of ICA components reveals parcellations within brain regions that may be relevant to claims about specificity and versatility. For example, a comparison of the spatial maps for element 13 (*tom, mental, story, social, and mentalizing*) and element 26 (*default, rest, restingstate*) yields considerable dissociation within the medial prefrontal cortex and bilateral temporoparietal junctions (Figure 6).

## “Periodic Table of Neural Elements”

(9) visual							(8) production
(48) audiovisual integration	(2) auditory	(3) <u>emotion</u> <u>neutral</u>	(13) <u>tom</u> <u>mental</u> <u>story</u> <u>social</u> <u>mentalizing</u>	(6) <u>sentence</u>	(58) strategy regulation	(34) tapping	(1) <u>motor</u> <u>hand</u> <u>movements</u> <u>finger</u>
(45) experience	(54) stimulation pain animal	(10) <u>reward</u>	(26) <u>default</u> <u>rest</u> <u>restingstate</u>	(4) <u>phonology</u> <u>word</u>	(43) rule planning	(65) difficulty	(7) <u>eye</u> <u>movements</u> <u>saccadic</u>
(47) shock physiological	(25) <u>somatosensory</u> <u>tactile</u> <u>stimulation</u>	(32) <u>negative</u> <u>positive</u>	(42) face facial social expression	(40) semantic	(38) difficulty correct perceptual demand feedback maintenance monitoring	(28) <u>nogo</u> <u>gonogo</u> <u>inhibition</u> <u>goal</u>	(14) <u>hand</u> <u>actions</u> <u>movements</u>
(64) videos	(20) motion visual	(51) positive negative	(41) personality personal	(35) readers reading	(52) maintenance	(12) arithmetic calculation	(19) <u>foot</u> <u>limb</u>
(5) <u>pain</u> <u>heat</u> <u>noxious</u>	(49) object spatial scene imagery visual mentalizing	(24) <u>taste</u> <u>rating</u> <u>food</u> <u>eating</u> <u>olfactory</u> <u>reward</u> <u>physiological</u>	(55) verbal verb	(62) verb verbal	(33) <u>incongruent</u> <u>congruent</u> <u>conflict</u> <u>stroop</u> <u>interference</u>	(16) work load working 2back memory executive 1back	(37) <u>motor</u> <u>muscle</u> <u>movements</u> <u>imagery</u> <u>somatosensory</u> <u>imagine</u>
(57) picture naming name	(56) pain sensation rating	(46) choice decision	(53) learning	(60) familiarity	(18) <u>autobiographical</u>	(23) <u>recollection</u>	(21) <u>remembered</u> <u>encoded</u> <u>remember</u>

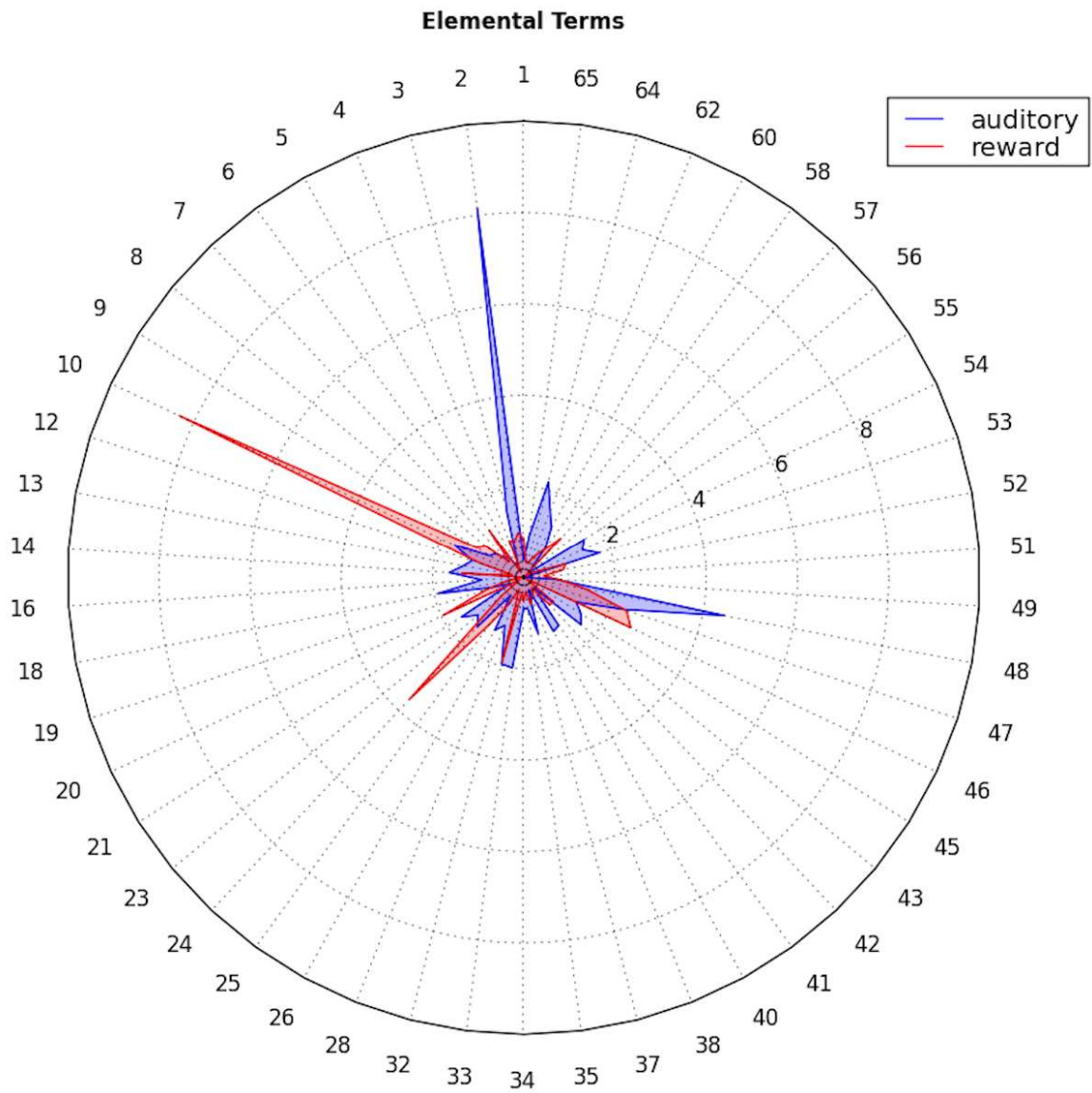
Figure 5: The “Periodic Table of Neural Elements.” From left to right, the neural elements are generally arranged by relevance to sensory input and motor output. Color categories were initially derived from the hierarchical clustering algorithm, such that underlined elements (or italicized for the green group which has two clusters) within the same color group had some degree of clustering according to the algorithm. The other elements belonging to each group were manually arranged into groups for visualization purposes. Numbers indicated for each element denote the numbering of the components from the most to least amount of variance explained from the original ICA decomposition.



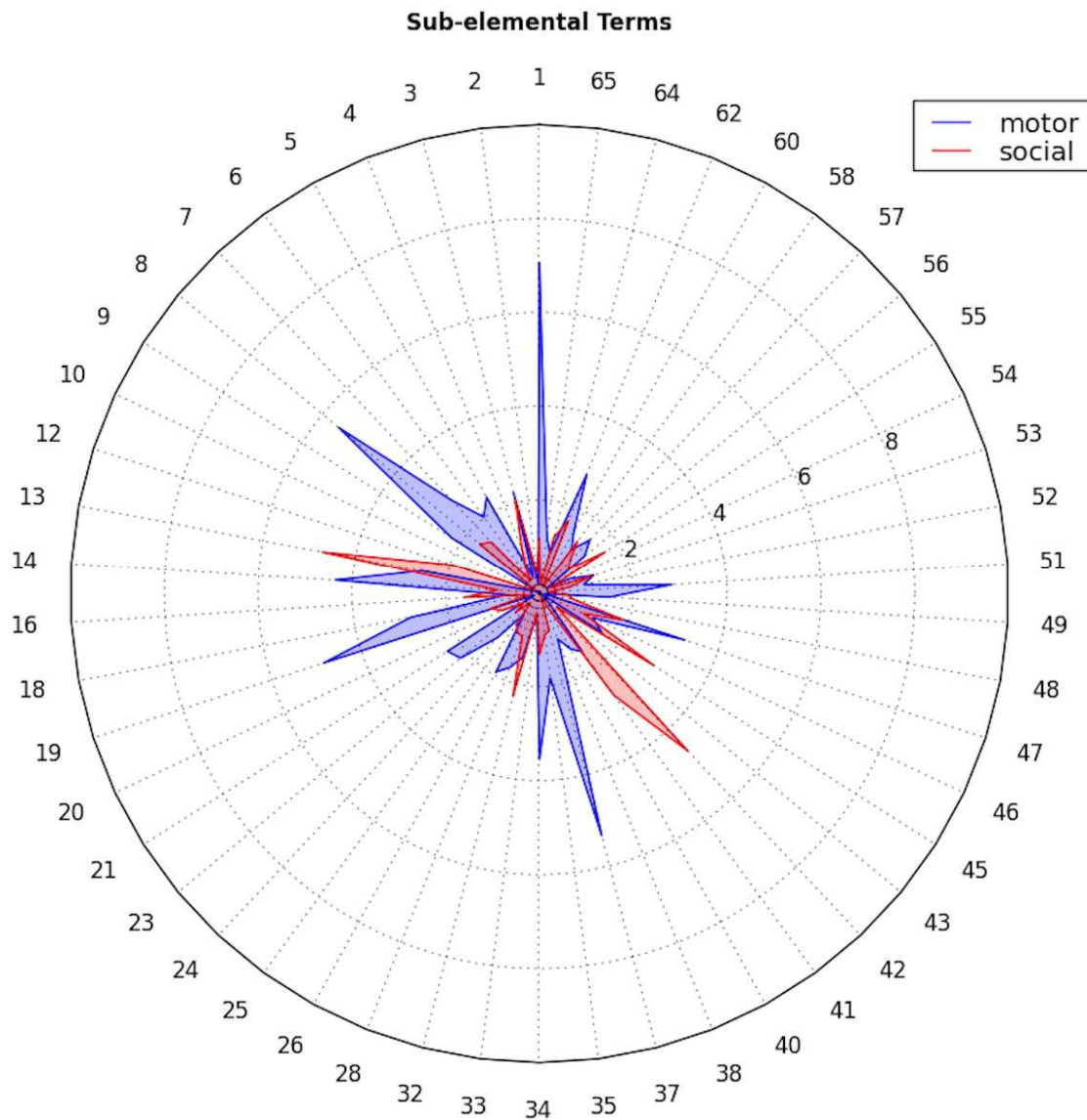
**Figure 6: Default mode and mentalizing neural elements have dissociable patterns of activation in temporoparietal junction, medial prefrontal cortex, and precuneus. Images are thresholded at  $z > 3$ . The maps are drawn from the spatial maps resulting from ICA decomposition of terms in the Neurosynth database.**

### 3.3.4 Term types

The terms in the Neurosynth database can themselves be categorized by their relevance to the components. Some terms will have a strong one-to-one mapping with a single neural element, such that the term will have a specifically strong relevance to one component and not any particularly strong relevance to any others. Two examples of such terms are *auditory* and *reward* (Figure 7). We will call these Neurosynth terms *elemental terms*. Other terms lack a specific one-to-one correspondence, but have a high association with multiple neural elements. Two examples of such terms are *motor* and *social* (Figure 8). We will call such terms *sub-elemental terms*.



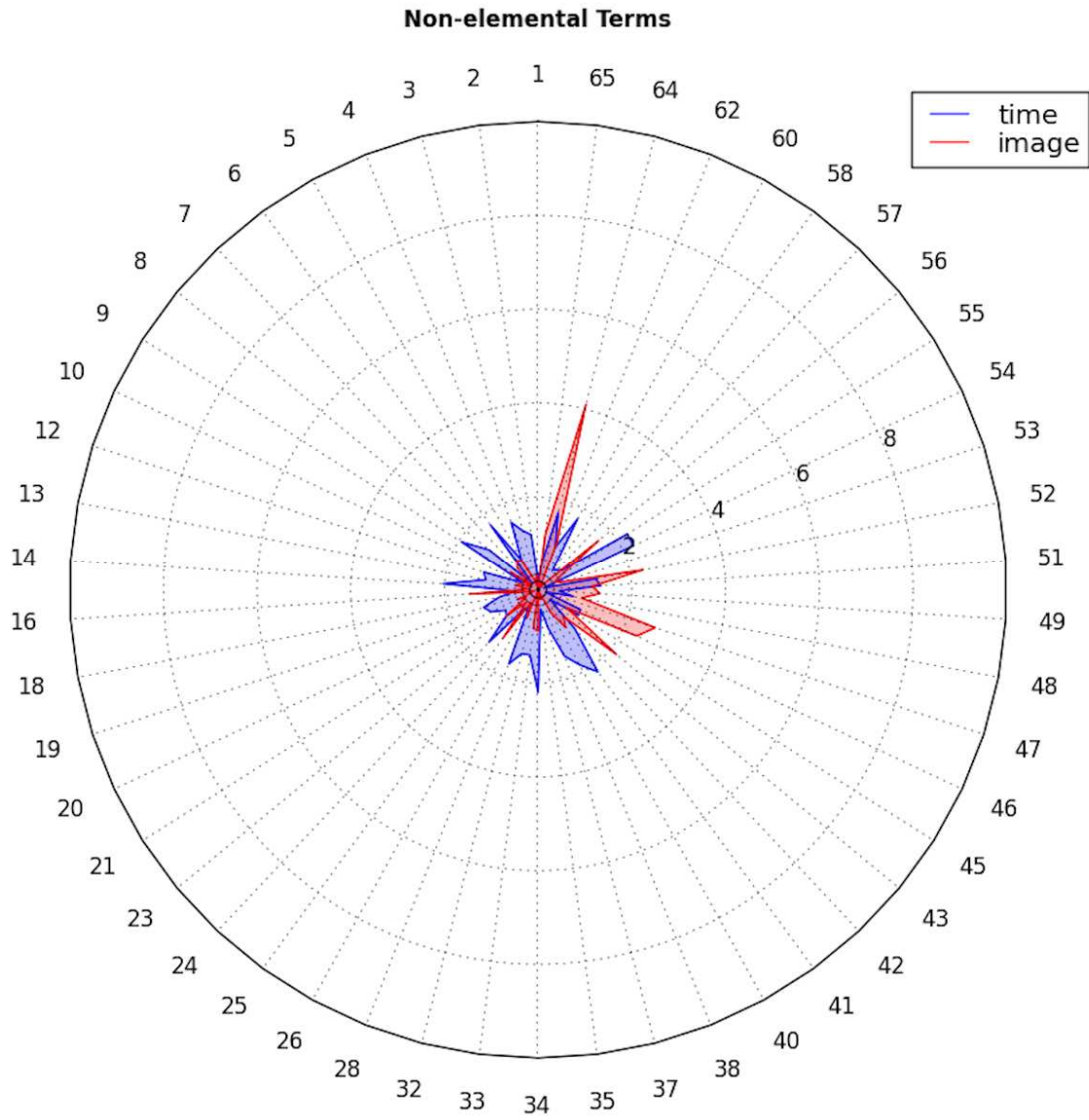
**Figure 7: Elemental terms have a strong one-to-one relevance to a particular neural element. In this case, “auditory” is highly associated with neural element 2, while “reward” strongly corresponds to neural element 10. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle.**



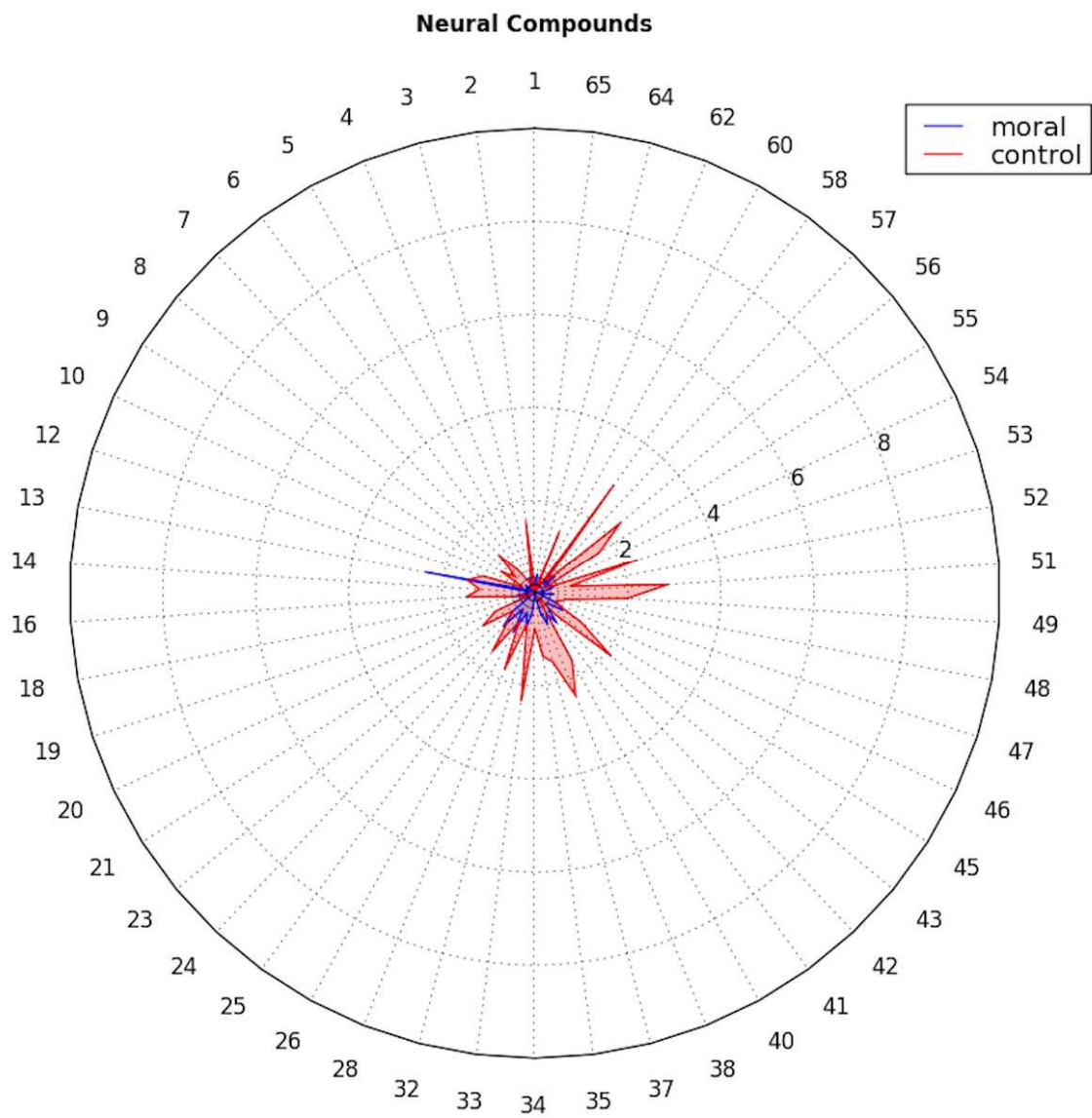
**Figure 8: Sub-elemental terms do not have a strong one-to-one mapping to any particular neural element, but rather have a high relevance for multiple neural elements. Here, “motor” and “social” have multiple strong peaks. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle.**

The members of a different class terms – *non-elemental* terms – have no particular one-to-one mapping with any particular neural element but also do not have any strong mappings. Two examples are *image* and *time*, which have a broad and low distribution across all neural elements. This is because such terms are used in many different contexts within the literature. A quick survey confirms the heterogeneity of the use of *time* in the literature, such that relevant studies include experiments directly about *time* (Wittmann, Simmons, Aron, & Paulus, 2010) and studies on wholly different topics (Aziz-Zadeh, Kaplan, & Iacoboni, 2009).

There is a distinct sub-group within the *non-elemental* terms that do not exhibit this type of semantic heterogeneity, which we will call *neural compounds*. These are terms that truly represent the combination of several different neural elements in unique proportions to give rise to neural processes with emergent characteristics. We propose that two such examples of such terms are *moral* and *control*. Their elemental compositions are displayed in Figure 10. As opposed to *time*, *moral* is only tagged in 31 studies, and all of them have something to do with directly studying the neural processes underlying moral cognition.



**Figure 9: Non-elemental terms do not have any particularly specific correspondence to any single neural elements and do not have moderate to low association across all elements. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle.**



**Figure 10: Neural compounds are composed of the combination of several different neural elements. Relevance measures from the “time courses” of the ICA decomposition are plotted as the radial distance from the center of the graph, and the neural elements are labeled along the circumference of the circle.**

### **3.4 Discussion**

Neurosynth currently represents the largest publicly available database of automated meta-analyses of fMRI studies (Yarkoni, Poldrack, & Nichols, 2011). One unique feature that sets it apart from other meta-analytic databases is the calculation and provision of reverse-inference maps that quantify the specificity of brain coordinates for certain psychological processes. Using ICA, we decompose this rich dataset into the fundamental *neural elements* that ultimately give rise to the tremendous diversity of neural processes studied within the neuroscientific literature.

#### **3.4.1 Theoretical Implications of Elemental Distinction**

We believe the structure of the decomposition described in this present work may have implications for a broad range of theoretical issues throughout cognitive neuroscience. For instance, there has been an area of research concerning the nature of the neural processing underlying the resting-state as manifest through the default mode network. Spreng et al. (2009) used an activation likelihood estimation (ALE) to confirm a high degree of similarity between four types of studies: autobiographical memory, navigation, theory of mind, and default mode network. Interestingly, each of these components fall under separate neural components in the current study, where element 18 is *autobiographical*, element 49 is *object, spatial, scene, imagery, visual, and mentalizing*, element 13 is *tom, mental, story, social, mentalizing*, and element 26 is *default, rest, restingstate*. We do not disagree with the authors' conclusion that there may be some

homology among these different types of neural processes. Whereas their study highlights the commonalities between these processes, the current data highlights the differences and dissociability of such processes. Indeed, default mode network seems to be something that is separate from theory of mind and mentalizing, and the spatial maps provided here could allow for great specificity in reverse inference and the use of *a priori* regions of interest.

### **3.4.2 Semantics in Relation to Different Types of Components**

There will always be some gap between what happens in the brain and the language that we use to describe it. However, the automated analyses of Neurosynth allow for great strides in characterizing the relation between the two. The identified *neural elements* represent terms that have a strong mapping between semantics and neural processes. The term, *auditory*, strongly characterizes element 2, such that all other terms in the database have much lower relevance, comparatively. The *neural sub-elements* represent a slightly weaker mapping between semantic and neural processing. Here, terms like *social* do not necessarily refer to any specific process in the brain. However, its use is crucial for the semantic description of multiple elements, including higher-level and more abstract processing of mental states represented in element 13 as well a lower-level class of processing associated with exposure to faces and expressions in element 42.

The semantic-neural process relation is yet another degree weaker for many *non-elemental* terms, which seems to have been captured by Neurosynth purely because of

the high frequency with which such terms are used in the text of fMRI studies. Except for a minority of studies, terms such as *image* and *time* are not the primary research interest of the fMRI studies included in the database under these terms. Though future efforts to filter such terms may be useful, our current methods were able to characterize the difference between these terms and more relevant terms such as those represented by the *elemental* and *sub-elemental* terms.

Finally, there are a group of terms which we hypothesize to have high semantic-neural correspondence, yet still do not have high loadings onto the neural elements. Much in the way in which chemical elements combine in various ways to give rise to chemical compounds, *neural compounds* represent the confluence of distinct neural system to give rise to processes with emergent properties.

### 3.4.3 Component Filtering

Several pre-processing steps were performed after ICA decomposition and the presentation of the neural elements in Figure 5. One was to separate the participant-related vs. task-related terms. Terms such as *autobiographical*, *stroop*, *moral*, and *reward* describe the nature of the task itself, while other terms such as *sex*, *sexual*, *women*, and *child* elucidate participant-related factors. Making this distinction is important because the brain maps associated with these two different sets of terms are fundamentally different. For the task-related terms, the brain maps reflect real neural networks underlying human cognition; For example, it is truly the case that the amygdala tends to

be more activated when participants engage in a task that is designed to study *emotion*. However, if one looks at the ICA components associated with participant-related terms, the interpretation is different. The activation of regions such as the thalamus and midbrain represented in ICA component 31 (*sex, sexual, women*) does not reflect neural processing associated with tasks involving such features, but rather involve the report of gender-related differences in neural activation on a variety of different tasks. Because of the diversity of tasks associated with such a difference, participant-related terms do not reflect any neural network involved in any single domain of human cognition.

This distinction is not made by the automated text analysis methods employed by Neurosynth. However, the data-driven ICA decomposition of Neurosynth was able to separate these two different types of terms by assigning the participant-related terms to three distinct components and the rest to a different group of components. This provides evidence for the ability of the ICA methodology to recognize and separate spatial components that are inherently different in source.

Seven artifactual components were also identified and discarded from further task-related analyses. Previous studies have similarly found artifactual components associated both with resting-state and meta-analytic data sources (Smith et al., 2009). For resting-state data, the sources of these components have physical correlates, such as head motion or acquisitional artifacts. For meta-analytic data, such sources are not

relevant, so various other interpretations have been provided. One is that several preprocessing steps used by the ICA algorithm within MELODIC (FMRIB, 2014) necessarily introduce artifactual components (Smith et al., 2009). In our study, the patterns of activation for all seven of our artifactual components were consistent, exhibiting a distinct pattern of activation generally localized to the gray-white matter junction across the brain. We speculate that this may reflect some bias within the literature in reporting white vs. gray matter activations.

Finally, we found that four networks that did not positively relate to any succinct set of Neurosynth terms. In fact, all of these were components were best described as the negation of a small group of terms that had extreme negative relevance values. For the purposes of this study, we were most interested in characterizing the structure of positive neural activation across various tasks, so these components were set aside. However, future work on the nature of these negative components could provide a more complete account of the underlying structure of neural processes.

#### **3.4.4 Neural Elements and Diversity**

Is there anything particularly special about the way in which we have defined a neural element, as reflected by the set of 50 neural elements? A factor that influences the number of components that result from our analyses is how we constrained the ICA decomposition dimensionality. The ICA algorithm can be forced to reduce the set of the data to either a greater or fewer dimensions, such that we would ultimately have a set of

components either greater or fewer in number. Forcing greater dimensionality would split some current elements into separate parts, while forcing lesser dimensionality would result in the merging of many of the current elements. Previous research has varied the dimensionality of the ICA reduction (Smith et al., 2009) and has found that such a decision merely influences the level of description and not overall structural claims.

Even in chemistry, some level of subjectivity is required in labeling entities as basic elements, since atoms are composed of protons, neutrons, and electrons, and each of those sub-atomic particles are composed of yet smaller entities. However, a great deal of benefit is derived from defining and studying matter at the intermediate level of the atom. We propose that similar benefit could be found in the level of analysis presented in this work, and this utility may arise from the fact that the dimensionality was chosen in a purely data-driven way.

Also like in chemistry, future research is bound to reveal novel elements. New tasks, methodologies, and research directions are likely to reveal that several other basic *neural elements* underlie cognition and that they can combine with one another and with the *elements* currently presented, to form *neural compounds*. New research will also better elucidate the nature of many of the elements presented here in more detail. For instance, the fact that *videos* is the best single-term descriptor of element 64 may point to the fact that researchers have not quite been able to match the semantic description of the

relevant psychological process to the underlying neural computation. The same may hold true for many other components, and the description of these components will be refined progressively as further research provides more data for semantic-neural mapping.

We propose that one immediate practical utility of our framework is through the provision of the most specific and independent set of ROIs based on the largest set of fMRI studies available. As mentioned above, the ICA decomposition has yielded maximally distinct spatial maps for seemingly similar neural networks such as mentalizing network and the default mode network. We make further use of this utility in Chapter 6 to address remaining concerns from Chapter 2.

A second practical utility is through the analyses of neural compounds. Several different literatures within cognitive neuroscience theorize about the neural compositions of various constructs, and we propose that the methodology presented in this chapter can be useful in providing a more data-driven approach to analyzing the contents of various psychological constructs. At this point, we turn back to the nature of morality and the ongoing debate regarding the nature and composition of moral cognition.

## 4. The Elemental Composition of Morality as a Neural Compound

### 4.1 Introduction

"Everything should be made as simple as possible, but not simpler." - Albert Einstein

The history of human thought is rife with attempts to boil morality down to its essence. Both descriptive and normative theories of morality have been proposed by philosophers, psychologists, and neuroscientists. Immanuel Kant held that morality was captured in a terse categorical imperative (Kant, 2002). Jeremy Bentham proposed that morality depended solely on maximizing happiness or "utility" (Bentham, 1907). More recently, empiricists have put forth distillations of how morality works in the mind/brain. In one prominent example, Gray et al., (2012) contend that mind perception ("mentalizing" in this dissertation) is the essence of morality. They propose a cognitive template - consisting of a moral dyad of an intentional agent and a suffering moral patient - represents the fundamental essence of morality. Their proposal is supported by a substantial body of behavioral and neuroimaging work that highlight the importance of mind perception in moral judgments.

Gray et al.'s simplification is attractive, especially because of its novel and elegant incorporation of neural data (particularly with the unique role right temporoparietal junction). But numerous counterexamples may serve as symptoms of oversimplification. Sinnott-Armstrong (2012) points out that Gray et al.'s proposal has trouble addressing prohibitions on incest, masturbation, and suicide (among others).

Baumeister and Vonasch (2012) propose that the essence of morality could also include self-regulation, free will, and culture. Monroe et al. (2012) contend that the main components of the "moral dyad" – intentionality, harm, and suffering" – are not necessary for moral judgment at all.

In fact, Gray et al. themselves point to a major weakness in their own theory. (2012). The theory has particularly difficult time explaining disgust. Previous work on disgust has conceptually separated it from other domains of morality (Graham et al., 2009), and neuroimaging studies have suggested that distinct neural networks underlie computations of disgust (Moll et al., 2005; Parkinson et al., 2011; Schaich Borg et al., 2008).

Leveraging meta-analytics neural data is one powerful approach to elucidating the structure of morality and disgust. We first propose the use of a novel, data-driven method of finding hierarchical ontological structures of neural processes. This structure will arise from a comprehensive survey of neural processes studied by neuroscientists and identifies two levels of process. The first consists of elemental neural processes, which compose the basic building blocks of human cognition. The second consists of neural compounds, which are formed from various combinations of elemental neural processes. We then use this ontology to explore the neural essence of morality.

There are three possibilities for the structure of morality. First, it may be the case that morality is itself an elemental neural process. Given its central importance to the

survival of humans and societies, it may have evolved as a unique module within the brain; it repeatedly performs the computations needed to cope within a highly social environment. If this is the case, mentalizing could not be the essence of morality since morality could not retain its essential qualities by being decomposed into simpler parts. Instead, if morality is found not to be an elemental process, then this could be consistent with a mentalizing essence for morality. However, if other elemental neural processes also seem integral to morality, this would suggest that the mentalization essence is an oversimplification that fails to sufficiently characterize morality. We conclude that morality is not elemental, and in addition to mentalizing, other components also seem important. Of these, one in particular seems to stand out: *taste*.

## **4.2 Methods**

The methods for ICA decomposition of the Neurosynth database and subsequent analyses can be found in greater detail in Chapter 3. In short, reverse inference brain maps were obtained from the Neurosynth Core Tools Package (Yarkoni, 2013). These maps were then submitted to spatial ICA decomposition using FSL (FMRIB, 2014). A hierarchical ontology was found identifying the basic neural elements of human cognition. We then explored several neural compounds in relation to their composition of neural elements. One important neural compound further discussed in this chapter corresponds to the unthresholded reverse inference map for *moral* taken from the

Neurosynth database (Yarkoni, 2013). It is based on data from 33 different studies, and the studies can be found on the Neurosynth website (Yarkoni et al., 2011).

To our knowledge, there has only been one direct meta-analysis of the neural networks involved specifically in moral cognition, which was performed by Bzdok et al. (2012). They calculated large-scale activation likelihood estimates (ALE) based on 247 handpicked neuroimaging studies, which incorporated data from 1790 participants. The Neurosynth map and the map from Bzdok et al. have very similar patterns of activation. Significant clusters of activation in both maps were found in the dorsomedial prefrontal cortex, ventromedial prefrontal cortex, frontopolar cortex, precuneus, right temporoparietal junction, left temporoparietal junction, right temporal pole, right middle temporal gyrus, and left amygdala. One difference was that the Neurosynth map additionally included significant activation in the right amygdala.

### **4.3 Results**

*Moral* does not seem to have a strong one-to-one mapping with any particular neural element (Figure 11). Indeed, morality was not identified as an elemental term from Figure 5. Although *moral* has an overall weak relevance to all neural elements, it does have a particularly strong relevance to element 13 (*tom, mental, story, social, mentalizing*). Additionally, a commonly discussed component of morality, *disgust*, itself was not found to be an elemental compound. However, element 24 (*taste, rating, food, eating, olfactory, reward, physiological*) was a major component of both *moral* and *disgust*.



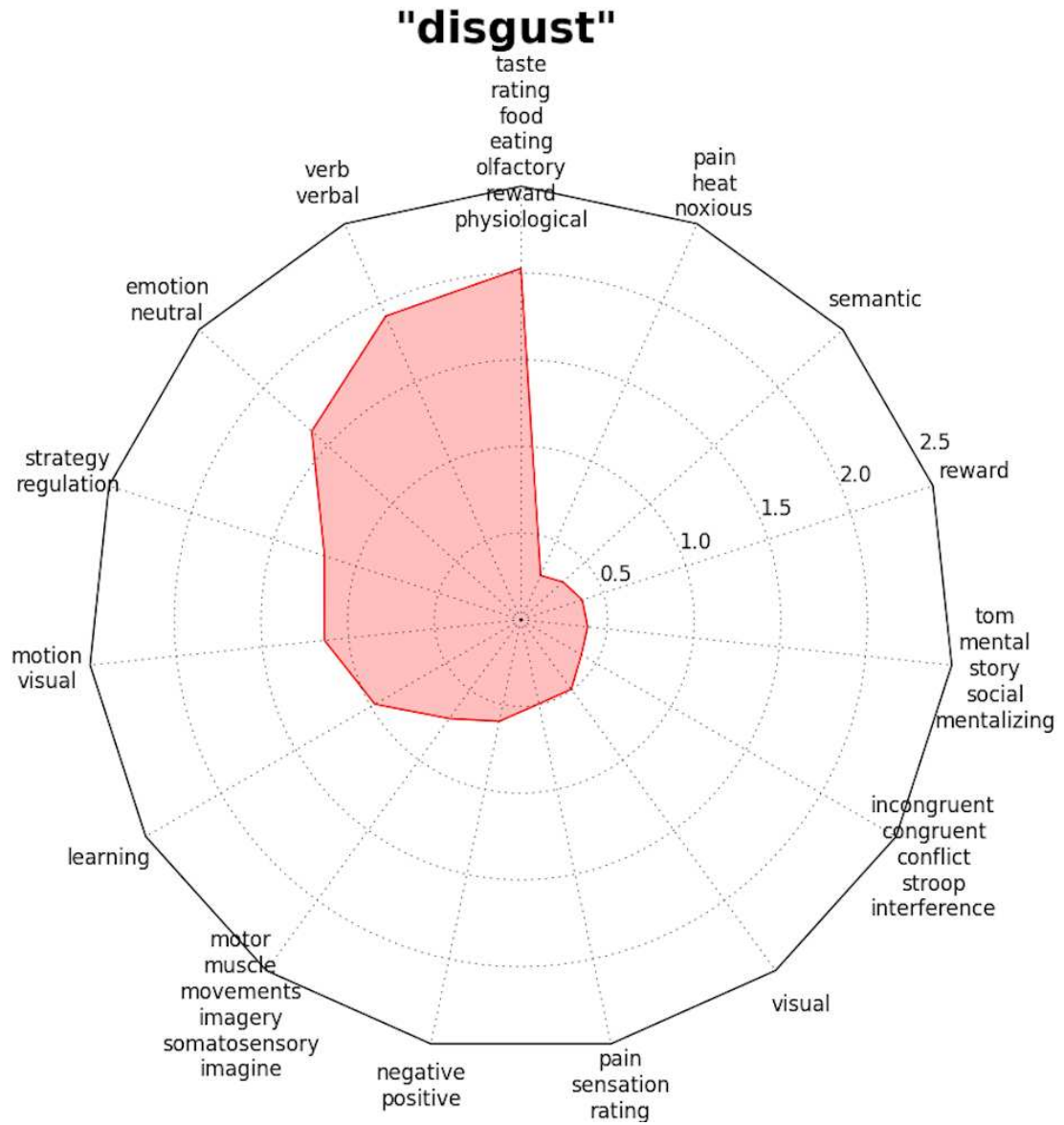


Figure 12: *Disgust*, as a common discussed component of morality, is not characterized as being a neural element. Its relevance is widely distributed among multiple neural elements. The methods used from Chapter 3 were applied, and 15 elements with the greatest relevance to *disgust* are listed here in descending order. The labels for these elements are taken from Figure 5.

## 4.4 Discussion

We applied the same methodologies developed in Chapter 3 to address examine the neural composition of moral cognition. We first find that *moral* is not an elemental neural process. Unlike *neural elements*, the neural circuitry of moral cognition is not a basic module that is repeatedly combined with other neural elements to form a diverse group of neural compounds. Instead, it seems to represent a *neural compound*, which is composed of a set of different *elements*.

The fact that *tom/mental/story/social/mentalizing* has the highest relevance to moral cognition seems to lend some support to Gray et al.'s (2012) claims. But does this mean that mentalizing is the *essence* of morality? Several other components seem also to be important in characterizing the neural processing underlying moral cognition. The *negative/positive* component mostly represents activation within the VMPFC, and many studies have explored the role that this regions plays in moral cognition (Greene, 2007; Koenigs et al., 2007; Moll, de Oliveira-Souza, Bramati, et al., 2002). The *strategy/regulation* components mostly represents activation within the L DLPFC, and the studies supporting Greene's dual-process theory (Greene et al., 2008, 2004, 2001; Greene, 2010) have much to say about this neural element's role in competing with emotion for processing moral judgment.

A less intuitive element that contributes to *moral* is the *default/rest/restingstate* element. Very little attention has been given to this element in the literature, though see

Greene and Haidt (2002) for a very brief discussion. This component highlights the care that must be taken to draw valid reverse inferences – activations in regions corresponding to the *default/rest/restingstate* element have been claimed to represent emotional processing (Greene et al., 2001), and a direct comparison of such activations with the *neural elements* in Chapter 3 could provide more precision and confidence for such reverse inferences.

From this set of elements described above, there seems to be significant overlap with the components of *disgust*. Indeed, *disgust* did not represent a neural element itself: instead, it seems that its relevance to various neural elements is widely distributed (Figure 12). This provides a new perspective on previous behavioral (Graham et al., 2009) and neural work (Moll et al., 2005; Parkinson et al., 2011; Schaich Borg et al., 2008) that have primarily regarded disgust as a more basic component. Some shared components between *moral and disgust* include *strategy/regulation* and *negative/positive*. Perhaps this is not surprising if disgust can be conceived of as a specific domain of morality, which engages many of the same processes as the *moral* as a whole.

But of all of the shared components, the most unique aspect of *disgust* features activation within the insula and orbitofrontal cortex and is represented by the *taste* element. That is, moral judgment in the *disgust* domain is distinct from the other domains of morality (e.g., harm and fairness) because of its much higher association with *taste* rather than *mentalizing*.

The importance of *taste* as a unique component suggests future avenues for study. Though theoretical connections association between *moral* and *disgust* has been heavily studied, perhaps more focus could turn to the association between *taste* and *moral*. A brief survey of the 34 studies on *taste* from the Neurosynth database (Yarkoni, 2011) does not yield any studies specifically about morality.

Extending beyond the Neurosynth database, one recent study does point towards future work that could be done in this general area. Researchers studied the Beauty-is-Good stereotype, which states that participants tend to assume that attractive people have more virtuous characters (Tsukiura & Cabeza, 2011). They found that aesthetic and moral judgments were associated with a common network of neural activity in the insula and orbitofrontal cortex. However, aesthetic judgment itself is a higher-level concept, which would probably represent a *neural compound* in our ontology. Further work could elucidate how even more basic neural circuits, such as those involved in the evaluation of olfactory and gustatory stimuli, give rise to moral judgments and decisions.

All of this provides challenges for the mentalizing-essence theory proposed by Gray et al. (2012). Their best defense is that disgust (or perhaps *taste* to be more precise) evolved as a module within the brain to protect humans from harm. It is a cognitive heuristic for motivating fast and reliable behaviors of avoidance. Because of its etiology, even moral judgments of disgust boil down to a perception of harm.

We propose that this is not a sufficient explanation, particularly if the *essence* of morality that the authors are referring to is a psychological/neural *essence* rather than an etiological essence. Concerning the latter, all of the *neural elements* presented in Chapter 3 presumably evolved from a whole host of evolutionary sources. For instance, the group of elements related to language would have evolved out of the advantages conferred by more complex communication. But finding an etiological essence does not seem to be the relevant enterprise for Gray et al.'s project or our project in this dissertation. The fact that separate modules exist in the brain, as supported by the fact that *taste* is a unique element from *mentalizing* in composing morality, suggests that one cannot simply reduce the former into the latter in neurological terms. Instead, the brain has evolved separate neural systems that seem to deal with harm and disgust in different ways, and the fact that both modules are adapted to help us avoid danger in general is irrelevant to the neural essence question. Accordingly, the *essence* of morality is not mentalizing.

But one could even go further to suggest that morality is not one unified construct at all. As neural evidence and theory suggest (Parkinson et al., 2011; Sinnott-Armstrong, 2012), perhaps different components of morality evolved to address threats in different ways at different times in evolutionary history, and therefore, would have different characteristics psychologically and neurally. Further empirical work and philosophical work may address whether some combination of the shared elements

discussed, like *negative/positive* and *strategy/regulation*, represent the unifying core of morality, while other varying degrees of other elements like *mentalizing* and *taste* determine the relevant domain of morality.

Finally, we propose that our data-driven methodology helps to address some biases that can arise from interpreting the literature as a whole. Though there have been previous meta-analyses, which include many more studies than this current one, the interests of the authors necessarily limit the scope of such work. Bzdok et al. (2012) focus on the roles of mentalizing, empathy, and control and their meta-analysis, but they neglect the role that disgust-related judgments may play in morality. To some degree, the same is true for attempts to simplify morality from Gray et al. (2012). Future work using methodologies similar to those in Chapter 3 and this chapter may provide further insight into morality as well as a large collection of other *neural compounds* within the cognitive neuroscience literature.

Throughout the last four chapters, we have surveyed several pieces of evidence elucidating the components that compose morality in the brain. The main idea is that these components interact with one another in unique ways and give rise to moral judgment in a bottom-up fashion. This has been the implicit working model within the moral neuroscience literature up to this point. However, an intriguing puzzle has garnered much attention in the fledgling field of experimental philosophy, called the

*Knobe Effect*. And this puzzle may warrant a radical revision to the current working model morality explored thus far.

## 5. Two Distinct Moral Mechanisms for Ascribing and Denying Intentionality

### 5.1 Introduction

Intentionality is foundational to many of our social interactions and institutions. For wrongful killing, it escalates a lesser charge of manslaughter to first-degree murder (*Murder in the first degree, N.Y State Penal Law Section § 125.27, n.d.*). In other legal contexts, its role in determining criminal culpability remains a primary focus in the most recent U.S. Supreme Court cases<sup>3</sup>. However, consider the following vignette from the nascent field of experimental philosophy (Alexander, 2012; Knobe & Nichols, 2008; Nichols, 2011):

The CEO knew the plan would *harm* the environment, but he did not care at all about the effect the plan would have on the environment. He started the plan solely to increase profits. Did the CEO intentionally *harm* the environment?

Most participants say “yes” (Feltz, 2007; Knobe, 2003, 2005, 2010), but consider a change in a single word:

The CEO knew the plan would *help* the environment, but he did not care at all about the effect the plan would have on the environment. He started the plan solely to increase profits. Did the CEO intentionally *help* the environment?

Here, most participants say “no” (Feltz, 2007; Knobe, 2003, 2005, 2010). Through vignettes like these, experimental philosophers have repeatedly shown that actions

---

<sup>3</sup> *Rosemond v. United States*: An ongoing Supreme Court case as of this writing where the core question being addressed is whether the offense of aiding and abetting requires proof of “simple knowledge” or “intentional facilitation.”

leading to negative consequences are judged to be more intentional than otherwise similar actions leading to positive consequences – often called the *Knobe Effect (KE)* (Feltz, 2007; Knobe, 2005, 2010). Importantly, this finding may be inconsistent with a long lineage of moral theories originating from Aquinas’s *Doctrine of Double Effect* from the Middle Ages (Aquinas, 1988) to contemporary theories of a universal moral grammar (Mikhail, 2007). According to these theories, intentionality serves solely as an input for moral judgments of blame and credit, rather than the reverse. There is controversy over whether the *KE* truly represents a violation of this assumption (Guglielmo & Malle, 2010; Knobe, 2010; Machery, 2008; Mallon, 2008; Nadelhoffer, 2006; Phelan & Sarkissian, 2008; Sripada, 2009; Utlich & Lombrozo, 2010; Wright & Bengson, 2009), and elucidating the mechanisms of the *KE* is integral to solving this debate. However, no significant consensus on a single theory or mechanism for the *KE* has arisen.

In experiment 1, we first rule out a broad range of possible one-process mechanisms. We created set of 40 novel scenarios – each with a negative and positive consequence variant – modeled after the original vignettes (Knobe, 2003). Participants (n=283) responded on a scale from 1 (Not Intentionally at All) to 8 (Completely Intentionally) in a self-paced task allowing for the measurement of response times. Previous work has highlighted the importance of individual differences in the *KE* (Cushman & Mele, 2008; Nichols & Ulatowski, 2007; Pinillos, Smith, Nair, Marchetto, &

Mun, 2011), so we sought to elucidate the mechanism of the *KE* through an extensive battery of individual difference measures drawn from personality psychology, decision making, and moral psychology.

In experiment 2, we tested the one-process hypothesis that emotional salience alone accounts for the asymmetry: negative consequences are judged as more intentional than positive consequences because the negative consequences are more emotionally salient. We presented participants from the online labor market Amazon Mechanical Turk (MTurk, n=386) with the original negative and positive conditions of the *KE* (Knobe, 2003). We also included a novel low-salience condition.

In experiment 3, we conducted an fMRI experiment (n=16) to demonstrate a double dissociation in mechanism across valences: emotion drives ascriptions of intentionality for negative consequences, while the denial of intentionality for positive consequences depends on the consideration of statistics norms. Participants were presented with the scenarios from Experiment 1 in an fMRI scanner (Figure 13A; n=16). In a post-scan session, participants were asked to rate the same vignettes regarding three other factors: *emotional reaction*, *statistical normativity*, and *moral judgment*.

## **5.2 Methods**

### **5.2.1 Experiment 1: Campus Behavioral Experiment**

#### **5.2.1.1 Participants**

Across four different rounds of experimentation (n=71, n=74, n=68, and n=70), 283 participants were recruited from Duke University and the surrounding community in all. All participants provided informed consent as part of an IRB exemption approved by the Institutional Review Board of Duke University.

#### **5.2.1.2 Task**

Participants completed self-paced and more verbose versions of a subset of 30 vignettes drawn from the comprehensive pool of vignettes presented below in “Experimental Stimuli: Vignettes.” Participants answered on a scale from 1 (Not Intentionally at All) to 8 (Completely Intentionally). The entire vignette and question was presented within one screen. The end-labels of this scale were counterbalanced across trials. Word count across negative and positive scenarios was balanced.

For each successive round of experimentation, we sought to identify whether various individual difference measures correlated with intentionality judgments. This included attempted replications of previously reported associated measures (Cokely & Feltz, 2009; Pinillos et al., 2011), as well as others drawn from several fields including personality psychology, decision making, and moral psychology. For the first round, participants subsequently completed scales for the Interpersonal Reactivity Index (IRI)

(Davis & Association, 1980), Tendency to Forgive Scale (TTF) (Brown, 2003), Machiavellianism Test (MACH-IV) (Christie & Geis, 1970), Vengeance Scale (Stuckless & Goranson, 1992), and Affective Intensity Measure (AIM) (Larsen, 1984). In the second round, participants subsequently responded to the IRI, Profile of Mood States (POMS) (McNair, Lorr, & Droppleman, 1971), Revised NEO Personality Inventory (NEO PI-R; only the extraversion subscale) (Costa & MacCrae, 1992), PAL (Personal Altruism Level) (Tankersley et al., 2007), and TTF. For the third round, participants completed the Cognitive Reflection Task (CRT) (Frederick, 2005), Rational-Experiential Inventory (REI-40) (Pacini & Epstein, 1999), Moral Foundations Questionnaire (MFQ) (Haidt & Joseph, 2004), AIM, POMS, IRI, TTF, NEO PI-R (extraversion), and BIS (Carver & White, 1994). For the fourth round, participants completed the NEO-ex, MFQ, and CRT.

### **5.2.1.3 Analysis**

No participants were excluded from analysis. We fit a hierarchical mixed-effects model (Snijders & Bosker, 2012) in order to account for the nesting of vignette trials within participants. These models also allowed the simultaneous examination of trial-varying and participant-varying effects. We fit models using SAS 9.3 Proc GLIMMIX (SAS, 2011) with adaptive Gaussian quadrature estimation (Rabe-Hesketh, Skrondal, & Pickles, 2002). The residual degrees of freedom were divided into between-participant and within-participant portions (Schluchter & Elashoff, 1990). Random intercepts were included, reflecting individual differences in participant means. Across experiments 1

and 3, random slopes provided little additional explanatory power and were excluded for parsimony. In each case, we analyzed null models (random intercept, no trial-level regressors) to estimate the intraclass correlation (participant-level variance/participant + trial level variance), reflecting the proportion of total variance accounted for by clustering of responses by participant. For all models, these values were substantively and statistically large, justifying a mixed model approach, as our measurements violated the independence assumption. Separate models were fit for the dependent variables intentionality rating (Supporting Equations 1, 3, and 4) and decision time (Supporting Equations 2, 3, and 5). Decision time residuals appeared log-normally distributed, so we fit a generalized linear mixed model using a log-normal distribution and an identity link function. The model equations are presented below, where  $i$  is trial,  $j$  is participant,  $r$  is the error,  $DT$  is decision time,  $trial$  is the number of trials the participant has previously seen,  $\gamma_{00}$  is the overall intercept, and  $u_{0j}$  is the random error component for deviation of the participant's intercept from the overall intercept.

#### Level 1

[Supporting Equation 1]:

$$Intentionality_{ij} = \beta_{0j} + \beta_1 val_{ij} + \beta_2 trial_{ij} + \beta_3 val_{ij} trial_{ij} + r_{ij}$$

[Supporting Equation 2]:

$$\ln(DT_{ij}) = \beta_{0j} + \beta_1 val_{ij} + \beta_2 trial_{ij} + \beta_3 val_{ij} trial_{ij} + r_{ij}$$

#### Level 2

For both dependent variables:

[Supporting Equation 3]:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Reduced form:

[Supporting Equation 4]:

$$Intentionality_{ij} = \gamma_{00} + \beta_{0j} + \beta_1 val_{ij} + \beta_4 trial_{ij} + \beta_7 val_{ij} trial_{ij} + u_{0j} + r_{ij}$$

[Supporting Equation 5]:

$$\ln(DT_{ij}) = \gamma_{00} + \beta_{0j} + \beta_1 val_{ij} + \beta_4 trial_{ij} + \beta_7 val_{ij} trial_{ij} + u_{0j} + r_{ij}$$

Correlations between scale measures and mean participant vignette responses were also analyzed.

## **5.2.2 Experiment 2: Online Behavioral Experiment**

### **5.2.2.1 Participants**

Through the online labor market, Amazon Mechanical Turk (AMT), 400 participants were recruited, redirected to Qualtrics, completed an online survey, and each paid \$0.25 in total. In all, 386 participants correctly answered an open-ended, comprehension catch question and were included in subsequent analyses. All participants provided informed consent as part of an IRB exemption approved by the Institutional Review Board of Duke University.

AMT has become an increasingly popular experimental tool and provides access to a large sample pool shown to be considerably more diverse than a typical American

college population. Though there is some debate about the validity of using AMT as a source for study (Chandler, Mueller, & Paolacci, 2013; Rand, 2012; Shapiro, Chandler, & Mueller, 2013), numerous replication studies have demonstrated that data collected on AMT is highly reliable and consistent with other methods of data collection (Buhrmester, Kwang, & Gosling, 2011; Chandler et al., 2013; Paolacci, Chandler, & Ipeirotis, 2010; Rand, Greene, & Nowak, 2012; Rand, 2012; Shapiro et al., 2013).

Several elements of our task design further ensure the quality of our data. As described in more detail below, we only included variations on one vignette instead of many to ensure that the task burden was not high (designed to be 2-3 minutes in duration). We also only included participants who had a previous approval rate of greater than 98% on AMT. There is the concern that with the growing popularity of AMT, participants may have repeatedly encountered similar tasks from various research groups, including our own (Chandler et al., 2013). We included a question directly asking whether he or she had seen a survey similar in nature to ours before, and we employed mechanisms within AMT and Qualtrics to prevent the same user or IP address from completing the survey more than once. Finally, we limited our participants to only those from the United States. Though there has been growing concern that users outside the United States have been able to circumvent these restrictions (Shapiro et al., 2013), the *KE* task has previously been shown to generalize well to other languages and cultures, particularly Hindi (Knobe & Burra, 2006).

### 5.2.2.2 Task

All participants read and completed negative and positive versions of Scenario #4 from “Experimental Stimuli: Vignettes,” provided below. This is the first and most commonly used *Knobe Effect* vignette in the literature (Knobe, 2003, 2010). Additionally, participants read a neutral condition vignette:

The chairman started a plan to increase revenue. He did not care at all about the effect the plan would have on the color of the product. He knew his plan would make the product yellow. Did the chairman intentionally change the color of the product?

Participants answered on a scale from 1 (Not at all intentional) to 8 (Completely Intentional). The entire vignette and question was presented within one screen. The end-labels of the scale were counterbalanced across scenarios and participants. Participants were also asked about the emotional salience of each of these vignettes: “How strongly did you emotionally react to the [environment/product color] being [harmed / helped / changed to yellow]? Participants answered on a scale from 0 (Not at all) to 10 (Extremely). The end-labels of the scale were also counterbalanced across scenarios and participants. Finally, participants were asked if they had ever seen these scenarios before and about the color of the product, as a catch question, from the neutral condition.

### 5.2.2.3 Analysis

Participants who did not correctly provide an answer to the catch question were excluded before data analysis. Ten participants reported that they had seen a scenario

that was similar to the one presented for this study. Exclusion of these participants did not significantly change the results of our analysis, so the presented analyses include all participants. Planned paired t-tests were performed across all three pairs of conditions. Single regression analyses between salience and intentionality ratings for each valence condition were performed in JMP 10.

### **5.2.3 Experiment 3: fMRI Experiment**

#### **5.2.3.1 Participants**

According to plan established before data analysis, twenty adults (mean age: 24, range: 18-32 years; 10 females) with normal or corrected-to-normal vision completed the study. We excluded four individuals from the final data analyses: one for an incidental anatomical finding, one for excessive head movement ( $> 2$  mm), and two for behavioral homogeneity precluding the inclusion of the parametric regressor in our GLM as described below. Prescreening excluded individuals with prior or current psychiatric or neurological illness. All participants provided written informed consent as part of a protocol approved by the Institutional Review Board of Duke University Medical Center.

#### **5.2.3.2 Stimuli and Tasks**

Participants completed four 20-trial runs of our Knobe Effect task (Figure 13A). Forty unique scenarios were constructed with a similar structure to those used in previous studies on the Knobe Effect, in which each scenario had two versions – one

where an agent's actions led to negative consequence and another leading to a positive consequence (Experimental Stimuli: Vignettes). To prevent participants from anticipating the moral valence of vignettes based upon task history, minor variations were incorporated into each of the scenarios across valence version, and the task was coded such that two versions of the same story did not appear in the same run in the fMRI scan. Vignettes were balanced for gender of agent and included both proper names of agents (e.g., "Bill") and general titles (e.g., "the doctor"; Experimental Stimuli: Vignettes). Participants were asked to answer on a scale from 1 to 8 with scale-end labels of "not intentionally at all" and "completely intentionally" randomized in left-right orientation trial-by-trial. There was a 2 second interval between trials. Different parts of the vignette were presented in isolation, as demonstrated in Figure 13A.

After the scanning session, participants were again shown the vignettes from the scanning session, but were prompted with three additional questions:

- How did the CEO's harming (or helping) the environment make you feel? [-3=Very Negative to 3=Very positive]
- How much blame (or credit) does the CEO deserve for harming (helping) the environment? [1=No Blame at All to 8=Extreme Blame]
- About how many people out of 100 in the general population would have harmed (helped) the environment under these circumstances? [0 to 100]

Results for question 1 were reverse coded to indicate the level of negative emotional reaction to the vignettes.

### 5.2.3.3 Behavioral Analysis: Hierarchical Mixed-Effects Model

Hierarchical mixed-effect models were fit in the same manner as those in Experiment 1.

The model equations for experiment 3 are presented below, where  $i$  is trial,  $j$  is participant,  $r$  is the error,  $trial$  is the number of trials the participant has previously seen,  $emot$  is the measure of *emotional reaction* (reverse-coded),  $stat$  is the measure of *statistical normativity*,  $\gamma_{00}$  is the overall intercept, and  $u_{0j}$  is the random error component for deviation of the participant's intercept from the overall intercept.

#### Level 1

[Supporting Equation 6]:

$$Intentionality_{ij} = \beta_{0j} + \beta_1 val_{ij} + \beta_2 emot_{ij} + \beta_3 stat_{ij} + \beta_4 trial_{ij} + \beta_5 emot_{ij} val_{ij} + \beta_6 stat_{ij} val_{ij} + \beta_7 trial_{ij} val_{ij} + r_{ij}$$

#### Level 2

[Supporting Equation 7]:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Reduced form:

[Supporting Equation 8]:

$$Intentionality_{ij} = \gamma_{00} + \beta_{0j} + \beta_1 val_{ij} + \beta_2 emot_{ij} + \beta_3 stat_{ij} + \beta_4 trial_{ij} + \beta_5 emot_{ij} val_{ij} + \beta_6 stat_{ij} val_{ij} + \beta_7 trial_{ij} val_{ij} + u_{0j} + r_{ij}$$

Unlike the analysis for experiment 3, decision times were not modeled and analyzed because the fMRI task restricted the pace by which participants could progress through and provide responses to the vignettes.

### 5.2.3.3 Behavioral Analysis: Mediation Models

We tested various mediation models (Figure 15) using trial-by-trial ratings for *emotional reaction*, *blame*, and *intentionality* for negative conditions and trial-by-trial ratings for *statistical normativity*, *credit*, and *intentionality* for positive conditions. The MBESS package for R (Kelley, n.d.; Preacher & Kelley, 2011) was used to calculate 95% confidence intervals with non-parametric bootstrapping (10,000 samples). Hypothesis testing was performed at  $\alpha=0.05$  by determining whether the bootstrapped 95% confidence interval was inclusive of 0 (Preacher & Hayes, 2004).

Further, we sought to test whether the “*emotional reaction* → *blame* → *intentionality*” and “*statistical normativity* → *credit* → *intentionality*” mediation models had significantly higher indirect effects than “*emotional reaction* → *intentionality* → *blame*” and “*statistical normativity* → *intentionality* → *credit*,” respectively. To do this, we took the difference in randomized bootstrap samples from the two theories compared and similarly performed hypothesis testing at  $\alpha=0.05$  by determining whether the 95% confidence interval of the difference in indirect effects was inclusive of 0. All mediations above were checked for interactions between the *X* variables and *M* (mediator) variables and possible moderator relationships were ruled out.

#### **5.2.3.4 fMRI Analysis: Image Acquisition and Pre-processing**

Functional MRI data were acquired using a 3T GE scanner with an 8-channel receiver using a spiral-in sensitivity encoding (SENSE) sequence. Four runs of 306 time points were acquired with TR=1.58, TE=30ms, voxel size=3.8mm x 3.8mm x 3.8mm, field of view=243mm, and flip angle=70°. Brain tissue was isolated using the brain extraction tool (Smith, 2002). The first six volumes of each analyzed run were discarded to account for magnetic stabilization. Differences in slice acquisition times were corrected using Fourier-space phase shifting. Spatial smoothing was performed with a Gaussian kernel with a full width at half maximum of 6mm. Grand mean scaling was performed across datasets from each run of each participant. A high-pass temporal filter was applied with a Gaussian-weighted least-squares straight line fitting with  $\delta = 100$  s. Functional images were registered to participants' high-resolution structural images with FLIRT, and subsequently, to MNI standard space with FNIRT (Jenkinson et al., 2002). Head motion was corrected by realigning the time series to the middle volume using FLIRT (Jenkinson et al., 2002).

#### **5.2.3.5 fMRI Analysis: Generalized Linear Model**

All fMRI analyses were performed with FEAT (fMRI Expert Analysis Tool) Version 5.98, which is part of FSL (FMRIB's Software Library). Time-series local autocorrelation correction was carried out with FILM (Woolrich et al., 2001). All presented analyses are taken from the "Knowledge" epoch when participants are first

able to determine the moral valence of the vignette. The first-level (within-run) analysis included two categorical regressors for moral valence of the consequence (positive vs. negative) and parametric regressors for normalized (within run and valence condition) participant ratings of intentionality. Participants providing consistent responses leading to rank deficient design matrices were excluded from the study (n=2). An additional categorical regressor was included for the “Question” epoch corresponding to whether the participant used the right or left hand in providing a response. This was used both as a nuisance regressor and an internal check of the validity of analyses based upon appropriate laterality of motor cortex activation. Additional nuisance regressors were included for the “Action” and “Attitude” epochs.

Second-level analyses (across-runs, within-participants) used a fixed-effects model, and third-level analyses (across-participants) used a mixed-effect model (FLAME 1).

#### **5.2.3.6 Neural Mediation Model**

Peak voxels were drawn from the Neurosynth reverse-inference map for the term “emotion.” Reverse inference maps indicate the specificity of relevant terms to specific brain coordinates and utilize a Bayesian statistic that controls for the number of studies associated with each term (Yarkoni et al., 2011). This yielded global peaks in the left (z-score = 17.4; MNI: (-20, -4, -16)) and right amygdala (z-score=14.2; MNI: (22, -2, -14)). Other brain regions usually associated with emotion, including the ventromedial prefrontal cortex, insula, medial orbitofrontal cortex, and anterior cingulate cortex, all

had markedly lower  $z$ -scores peaks (all  $z$ -scores  $< 7.0$ ). This suggested that of all these regions, amygdala seems to be most specific for emotional processing. Spheres with 8 mm radii were drawn at these coordinates in left and right amygdala. These spheres served as ROIs for finding voxels that significantly correlated on a between-participants basis with intentionality ratings. After small volume correction within this ROI, the activity from significantly correlated voxels was used for independent mediation analyses described below.

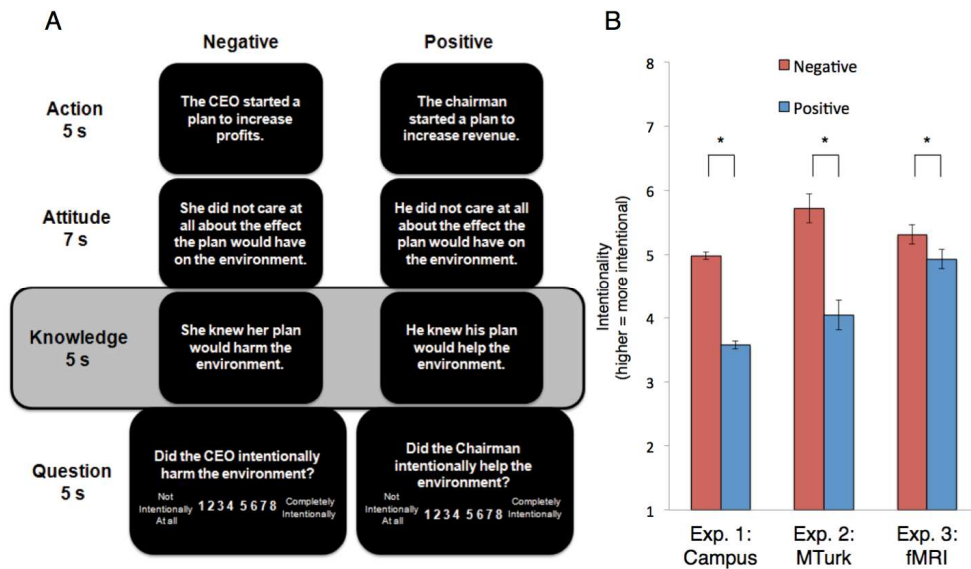
Using the MBESS package for R (Kelley, n.d.; Preacher & Kelley, 2011), we tested whether post-scan measures of *emotional reaction* to negative scenarios significantly mediated the relationship between activity in bilateral dorsal amygdala and intentionality rating. As negative controls, we also tested whether the same relationship held for positive consequences. The ROI analysis for positive consequences did not yield any significant voxels after small volume correction, so we obtained activation estimates from the same voxel coordinates that had been found in the negative consequence mediation analysis. We also did a whole-brain analysis for voxels whose activities significantly correlated between-participants with mean intentionality ratings. This analysis yielded activation in left dorsolateral prefrontal cortex (L DLPFC) as well as several occipital regions. We used the significant cluster of activation from L DLPFC as another negative control for a mediation analysis. On all the above mediation models, hypothesis testing was performed at  $\alpha=0.05$  by testing whether 95% confidence intervals

of the indirect effect estimates were inclusive of 0 (Preacher & Hayes, 2004). All mediations above were checked for interactions between the X variables and M (mediator) variables and possible moderator relationships were ruled out.

## **5.3 Results**

### **5.3.1 Experiment 1**

A hierarchical, mixed-effects model confirmed an asymmetry of intentionality in our novel set of scenarios: ratings for negative conditions were higher than those for positive conditions ( $\beta=1.64$ ,  $t(282)=20.2$ ,  $p<0.0001$ ; Figure 13), consistent with previous studies (Feltz, 2007; Knobe, 2003, 2005, 2010). Given the repetition and length of our task, we find evidence of practice effects in that there was a significant valence  $\times$  trial number interaction ( $\beta=-0.01$ ,  $t(8049)=-3.15$ ,  $p=0.002$ ): intentionality ratings for positive conditions significantly increased over successive trials while those for negative conditions significantly decreased. In a separate but analogous model with response time as the dependent variable, we found significantly longer decision times for positive conditions compared to those for negative conditions ( $\beta=2.00$ ,  $t(1209)=4.14$ ,  $p=0.0001$ ).

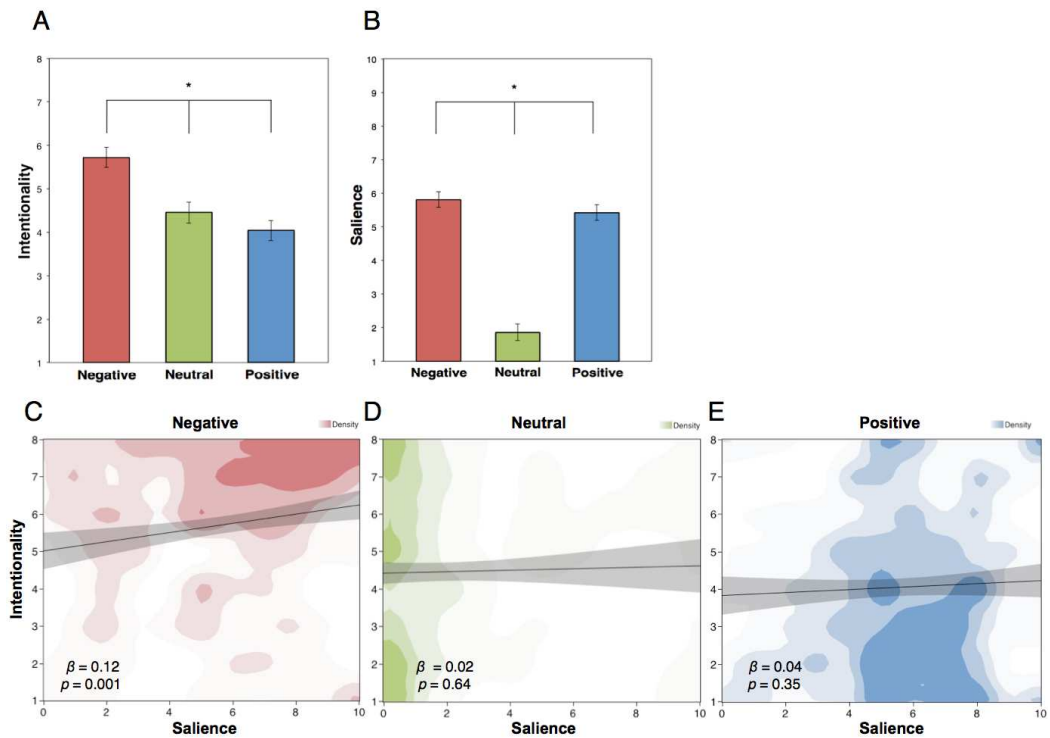


**Figure 13: Asymmetries in intentionality are robust across three different methods of experimentation. (A)** In the fMRI version of the task, participants read and responded to two versions of each general story. These versions differed in whether the agent’s actions lead to morally negative or positive consequences (40 pairs for 80 vignettes total). Participants provided ratings of intentionality on a scale from 1 (completely unintentionally) to 8 (completely intentionally), and the direction of the scale was counterbalanced trial-by-trial. Reported imaging results are derived from data collected during the “Knowledge” epoch. The ITI was 2 s. **(B)** At the group level, participants consistently rated actions in negative conditions as being more intentional than those in positive conditions across three different experiments. Model-free means are presented along with 95% confidence intervals for comparison across three different experimental designs. \*Indicates that the means are different according to paired *t*-tests for experiment 1 and according to hierarchical, mixed-effect modeling for experiments 2 and 3.

Numerous individual difference measures were collected following the task, including items such as the Interpersonal Reactivity Index (IRI) (Davis & Association, 1980) and the NEO Personality Inventory (NEO-PI-R) (Costa & MacCrae, 1992). We found no significant correlations between any of these scales or their component subscales with participants' mean difference between ratings for negative and positive consequences. Indeed, finding evidence of a single-process mechanism underlying the asymmetry remained elusive. However, analyzing ratings for negative and positive consequences separately yielded one suggestive result. There was a correlation between ratings for negative consequences and measures of moral harm sensitivity on the Moral Foundations Questionnaire (Graham et al., 2009) across two separate rounds of data collection ( $R = 0.23$ ,  $p=0.05$ ,  $n=68$  and  $R=0.30$ ,  $p=0.01$ ,  $n=70$ ). This is consistent with the fact that all of the negative consequences in our scenario set have to do with "harm" rather than any of the other moral foundations such as "purity" or "authority." Though limited, this data suggests that (1) moral judgment seems to have a specific role in producing the asymmetry and (2) separating the negative and positive components of the asymmetry may be fruitful, perhaps due to the lack of a single mechanism explaining the entire asymmetry.

### 5.3.2 Experiment 2

The aim of experiment 2 was to test the one-process hypothesis that emotional salience accounts for the asymmetry. Seemingly supportive of the emotional salience model, we found salience ratings to be higher for negative conditions than for positive conditions (Figure 14; paired  $t(192) = 2.28, p = 0.02$ ). However, data from a novel, low-salience and neutral condition contradict such a model. The emotional salience model predicted that the low-salience condition would have the lowest rating of intentionality of all. Instead, participants ascribed lower intentionality to positive compared to neutral conditions (paired  $t(192) = 2.58, p < 0.01$ ; Figure 13) even though salience ratings were much higher for positive compared to neutral conditions (paired  $t(192) = 17.8, p < 0.0001$ ).



**Figure 14: Emotional salience does not account for differences in intentionality ratings between outcomes with different emotional valence, but it does predict intentionality for negative outcomes. Participants (n=386) on AMT were presented three versions of scenario #4 differing in valence. (A) All pairwise comparisons among the three conditions were significantly different from one another. Participants ascribed higher intentionality for negative compared to positive (paired  $t(192) = 11.3$ ,  $p < 0.0001$ ), higher for negative compared to neutral (paired  $t(192) = 8.19$ ,  $p < 0.0001$ ) and lower intentionality to positive compared to neutral (paired  $t(192) = 2.58$ ,  $p < 0.01$ ). The data from negative and positive conditions were also presented in Figure 13B. (B) The neutral condition had significantly lower ratings of saliency than negative (paired  $t(192) = 18.03$ ,  $p < 0.0001$ ) and positive conditions (paired  $t(192) = -17.8$ ,  $p < 0.0001$ ). Negative conditions did have higher saliency ratings than those for positive conditions (paired  $t(192) = 2.28$ ,  $p = 0.02$ ). Error bars indicate 95% confidence interval. (C) For negative conditions, saliency ratings were positively correlated with those for intentionality. The same was not found in neutral conditions (D) or in positive conditions (E). Density plots are overlaid with a regression line with 95% confidence interval. \*All pairwise comparisons are significantly different from one another according to a paired t-test.**

These results bring focus to the positive condition, which has received little attention within the literature. It has implicitly been assumed that the positive condition was the relevant control condition for examining possible performance errors being made by participants in negative conditions (Nadelhoffer, 2006). However, the fact the positive condition has a lower intentionality rating than the neutral condition suggests that a unique process (independent of the process in negative conditions) may cause participants to deny intentionality in certain cases. Further analysis of the role of salience is consistent with this hypothesis. When analyzing the conditions independently as suggested by the data in experiment 1, we found an effect of emotional salience in predicting ratings of intentionality in negative condition ( $\beta=0.12$ ,  $t(384)=3.24$ ,  $p=0.001$ ), but not for neutral ( $\beta=0.02$ ,  $t(384)=0.47$ ,  $p=0.64$ ) or positive conditions ( $\beta=0.04$ ,  $t(384)=0.93$ ,  $p=0.35$ ; Figure 14). Emotional salience is an underlying mechanism for the negative condition but not others. However, there are still challenges to this interpretation: self-reports of emotional salience may not be ideal, and the mechanism for underlying positive consequences remains unknown. Further, even though the role of emotion has been implicated, it is unclear what role this has in determining whether morality can influence intentionality judgments. Experiment 3 was designed to address these issues by incorporating converging behavioral and neural evidence.

### 5.3.3 Experiment 3

Behaviorally, we fit hierarchical, mixed-effects model in a similar fashion to experiment 1. We again found a robust asymmetry: intentionality ratings for negative conditions were higher than those for positive conditions ( $\beta=1.95$ ,  $t(15)=4.76$ ,  $p=0.0003$ ). However, we also included *emotional reaction* and *statistical normativity* as predictors in the model. This allowed for the identification of a double dissociation in the mechanisms for negative and positive conditions. There was a significant valence  $\times$  emotional reaction interaction ( $\beta=0.46$ ,  $t(1209)=4.35$ ,  $p<0.0001$ ) as well as a significant valence  $\times$  statistical normativity interaction ( $\beta=-0.23$ ,  $t(1209)= -4.54$ ,  $p<0.0001$ ). Interrogation of these interactions yielded a significant effect for emotional reaction in negative ( $\beta=0.44$ ,  $t(1209)=6.30$ ,  $p<0.0001$ ) but not in positive conditions ( $\beta=-0.01$ ,  $t(1209)=-0.18$ ,  $p=0.85$ ) and a significant effect for statistical normativity in positive ( $\beta=0.20$ ,  $t(1209)= 5.33$ ,  $p<0.0001$ ) but not in negative conditions ( $\beta=-0.02$ ,  $t(1209)=-0.82$ ,  $p=0.42$ ; Figure 15). More negative emotional reaction ratings predicted higher intentionality ratings in negative conditions while numerically smaller assessments of statistical normativity (more rare) predicted lower intentionality ratings in positive conditions.

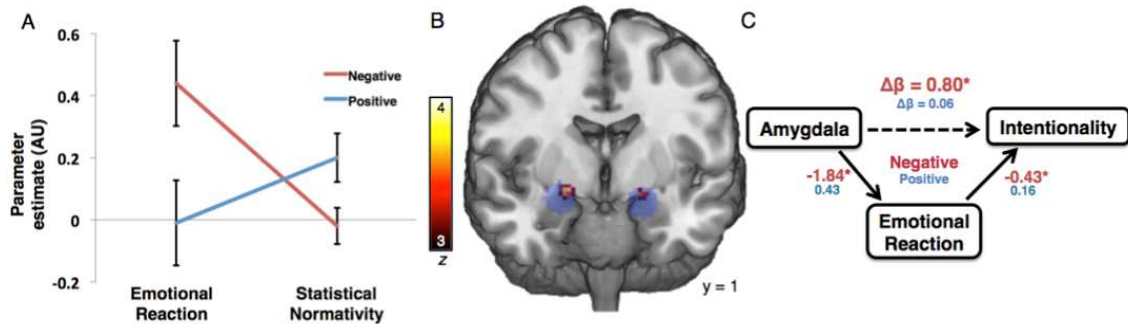


Figure 15: Converging behavioral and neural evidence suggests that *Ascription* leads to higher intentionality through an emotional mechanism while *Denial* leads to lower intentionality and is dependent on *statistical normativity*. (A) Behaviorally, *emotional reaction* significantly predicts intentionality ratings for negative conditions but not for positive conditions. Conversely, *statistical normativity* predicts intentionality ratings for positive conditions but not for negative conditions. The parameter estimates and 95% confidence intervals are presented from the hierarchical, mixed-effects model. (B) Activation in bilateral dorsal amygdala (red-yellow color map) was found to be positively associated with intentionality ratings for negative outcomes within ROIs identified from reverse inference maps of “emotion” from Neurosynth, indicated in blue. (C) This relationship was partially mediated by reports of emotion for negative consequences (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.80; 95% confidence interval = [0.07, 2.02]; Online Methods) while reports of positive emotion did not have a mediating role.  $\beta$  for separate negative and positive consequence mediation models are indicated, while the  $\Delta\beta$  indicates the change in beta value for the direct path after controlling for the indirect path.

We then tested for corroborating neural evidence for emotion in negative consequence trials by incorporating fMRI data into several causal models. We hypothesized that individual differences in activation in brain regions specific to emotion in negative conditions should lead subjects to more negative *emotional reactions*, which should in turn lead to higher ratings of intentionality. To identify relevant brain regions, we extracted BOLD response estimates from 8 mm spheres centered around peak voxels of the reverse-inference map for the term “emotion” from Neurosynth, a meta-analytic database that includes more than 5,800 fMRI studies (Yarkoni et al., 2011). These peaks, found in bilateral dorsal amygdala (Figure 14b), represent voxels most specifically associated with the term “emotion” compared to any other relevant term (Yarkoni et al., 2011; Yarkoni, 2011). A significant mediation model supported our hypothesis (19; Figure 14C), consistent with the role of amygdala activation and *emotional reaction* in negative consequences. Because of potential confounds in mediation analyses (Bullock, Green, & Ha, 2010), we ran several additional negative controls to help address these concerns. An analogous mediation analysis in the amygdala for positive conditions was not significant (Figure 15C). We also ran a whole-brain search for regions associated with intentionality in negative conditions and identified a cluster in the left dorsolateral prefrontal cortex. Here, we failed to find a significant mediation

model for signal extracted from DLPFDC and ratings of intentionality by *emotional reaction*.

Several other findings from the current experiment converge with results found in experiment 1. A direct contrast of positive and negative conditions heightened activation in distributed patterns of brain regions including lateral prefrontal cortex. This is consistent with the finding that participants had longer reaction times for positive compared to negative consequences and with the fact that judgments of intentionality for positive scenarios are driven by a non-emotional process dependent on statistical normativity. Additionally, we replicate the significant vignette valence  $\times$  trial number interaction ( $\beta=-0.01$ ,  $t(1209)=-2.31$ ,  $p=0.02$ ). Together, these trial effects are also consistent with a two-process model. If only one process were responsible for pushing ratings in negative consequences higher compared to those for positive, practice effects would only decrease over time in negative consequences. However, both negative and positive seem to move toward an implicit baseline.

Finally, we sought to address whether *moral judgments* can serve as inputs for judgments of intentionality vs. the reverse. Using the behavioral data, we analyzed various mediation models that differed in whether *moral judgment* or *intentionality* served as the mediator for a relationship with *emotional reaction* and *statistical normativity* for negative and positive conditions, respectively. We found *moral judgment* of blame to significantly mediate the relationship between *emotion reaction* and *intentionality* for

negative conditions (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.30; 95% confidence interval = [0.18, 0.43]) and *moral judgment* of credit to significantly mediate the relationship between *statistical normativity* and *intentionality* for positive conditions (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.10; 95% confidence interval = [0.05, 0.13]; Figure 16). Though significant, the models using *intentionality* as the mediator yielded significantly smaller mediating effects for both conditions.

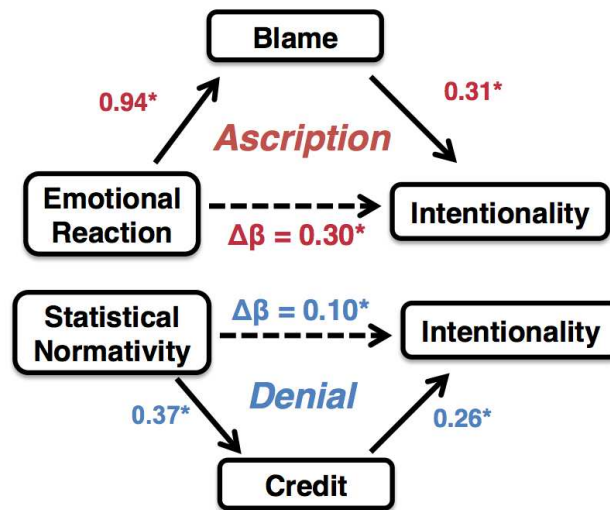


Figure 16: Moral judgments of blame and credit serve as inputs for intentionality ascription in both *Ascription* and *Denial*. Moral judgment of blame served as a significant mediator of the relationship between *emotional reaction* and *intentionality* in negative conditions (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.30; 95% confidence interval = [0.18, 0.43]). Moral judgment of credit served as a significant mediator of the relationship between *statistical normativity* and *intentionality* in positive conditions (Indirect Effect Estimate ( $\Delta\beta$ ) = 0.10; 95% confidence interval = [0.05, 0.13]).

## 5.4 Discussion

Across a series of three experiments, converging behavioral and neural evidence demonstrate two distinct and dissociable processes. Emotion drives higher ascriptions of intentionality for negative consequences, while statistical norms underlie the denial of intentionality for positive consequences. Further analysis shows that moral judgments of blame and credit can serve as inputs for intentionality judgments, rather than only the other way around.

For negative consequences, our data is consistent with theories describing the *KE* as an emotionally-based process of motivational bias (Alicke, 2008; Malle & Nelson, 2003; Nadelhoffer, 2006), but also addresses a major criticism of such theories heretofore: the lack of any positive evidence (Knobe, 2010). A second criticism is also addressed, coming from a previous study that showed an intact *KE* in a population of patients with blunted affect due to ventromedial prefrontal cortex (VMPFC) lesions (Young, Cushman, Adolphs, Tranel, & Hauser, 2006). Rather than the VMFPC, we demonstrate that the crucial emotional signals that lead to *Ascription* may be generated in the amygdala.

Within the literature, our conclusions regarding positive consequences and its underlying mechanisms represent a new focus. Just like in negative conditions, a motivational bias seems to drive intentionality judgments because participants view the

agents with a generally negative attitude. This attitude arises from the fact that the agent approaches doing something good, such as helping the environment, with indifference. This is supported by supplemental survey data: we found that 76% of participants had some sort of negative attitude towards the CEO when given only the first portion of the positive consequence scenario.

In contrast to negative consequences, participants here are motivated to withhold credit and intentionality from the agent, and such a motivation is not driven by emotion; instead, statistical norms play the major role. Statistical norms have previously been implicated in judgments of causality (Sytsma, Livengood, & Rose, 2012), and future work may elucidate whether causality has a mediating role in judgments of intentionality.

More broadly, we demonstrate a mechanism by which judgments of intentionality can be influenced by moral judgments. Previous work has integrated behavioral and neural data to study the mechanisms of moral judgment and punishment (Buckholtz et al., 2008; Greene et al., 2001; Koenigs et al., 2007; Moll, de Oliveira-Souza, Eslinger, et al., 2002; Schaich Borg et al., 2006), and more specifically, the influence of intentions on moral judgment (Kliemann, Young, Scholz, & Saxe, 2008; Koster-Hale et al., 2013; Mikhail, 2007; Moran et al., 2011; Young et al., 2007; Young & Saxe, 2009). We extend this literature by developing a novel conceptualization for intentionality judgment and its neural mechanisms, and we utilize this framework to identify how the

commonly conceived directionality between intentions and moral judgments can be reversed. A revised model of intentionality judgment, arising from this and previous interactions between philosophy and empirical experimentation, can have direct implications for the legal system where questions of intentionality remain foundational.

Beyond the legal system, the moral mechanisms for the *KE* also have important implications for a central principle in moral theory and practice, the doctrine of double effect (DDE). The doctrine asserts that it is morally wrong to cause harm intentionally in circumstances where it would not be moral wrong to cause harm unintentionally (Woodward, 2001). The DDE thus places intentionality ascriptions at the very foundation of moral reasoning. This principle was suggested by St. Augustine (1947) and St. Thomas Aquinas (1988) in the Middle Ages and since then has remained central to Catholic moral teachings as well as to many secular theories in moral philosophy (Kant, 2002) and moral psychology (Mikhail, 2007). In recent years, the DDE has been cited in arguments against terror bombing (White, 1985), against nuclear retaliation on cities during the Cold War (Norman, 1995), against some forms of contraception and abortion (Posner, 1992), and against active euthanasia and assisted suicide (Sanbar, 2004)—all on the grounds that these practices involve causing death intentionally. However, if ascriptions of intentionality already presuppose a prior moral judgment about the value of consequences, as the *KE* suggests, then the DDE would be threatened with circularity, showing that it cannot be fundamental in moral theory. The moral

mechanisms of Knobe effect could thus force us to rethink a large portion of moral theory in theology and philosophy.

## **6. General Discussion**

### **6.1 Overview**

The neuroscience of moral cognition has featured many distinct approaches and perspectives. Some researchers have focused on the emotional contributors to moral judgment, similar in approach to predecessors from philosophy and psychology. Others are focused on similarities that moral judgments have to those made in economic domains, particularly as they relate to the neural computation of value. A third approach is through the focus on the perception of minds, where some researchers have gone as far as to claim that mentalizing represents the true essence of morality. We unify two of these perspectives in the second chapter, where we analyzed a complex clinical decision-making task that employed two components of moral cognition: mentalizing and value. Mentalizing regions tracked the magnitude of the conflict of interest, and value regions tracked the degree to which participants modified their behavior in light of a policy of disclosure. Moving beyond the informal compositional model of morality in the literature, we laid the groundwork for formalizing such a model in chapter 3. There, an ICA decomposition of the largest available meta-analytic neuroimaging database yielded 50 distinct neural elements that form the foundation of human cognition. The same framework was brought to bear on the question of the composition of morality, where we found an important role for mentalizing. However, there was more to the story than this, especially as represented by an element localized in the

insula and the orbitofrontal cortex associated with *taste*. Finally, we found that even this formalized compositional model lacked the ability to explain an intriguing yet important phenomenon: indeed, moral judgments can paradoxically influence one of its components in a top-down fashion.

## **6.2 What is the essence of morality?**

Contrary to the various attempts throughout history to define or characterize morality in one simple description, I support the view that morality is not reducible to any single essence. In chapter 4, we demonstrated that *moral* was not an elemental term itself: it is not a building block the brain uses for synthesizing more complicated neural processes. Instead, we found it to be the combination of several neural elements. To take the chemistry analogy further from Chapter 3, *moral* represents a neural compound in very much the same way that water represents a chemical compound. Accordingly, one could ask, "What is the essence of water?" Perhaps one could propose that the true essence of water is hydrogen, since a water molecule has twice as many hydrogen atoms as oxygen atoms. Alternatively, one could propose that the essence of the water molecule is oxygen because it composes the vast majority of the molecule's mass. Either of these proposals could be reasonable depending upon the relevant essence is: the first would be appropriate for a numerical essence, while the second would be appropriate for a mass-related essence.

However, the main utility of defining compounds, whether neural or chemical, is based on the compound's emergent properties, and these properties only arise from the combination of more basic elements. In the case of water, its most remarkable properties arise from the emergent property of polarity, which results from the interaction between oxygen and hydrogen atoms that highly differ in electronegativity. This allows water to be a "universal solvent" in its liquid phase, and it is this property that allows water to be the foundation for all life. So if there is an answer to the question of what the essence of morality is, that essence must be able to fully explain all that makes morality important to us. Morality's importance, however, does not seem to be restricted to any single domain: it seems to span a wide range of phenomena including acts of murder, incest, and deception. Behaviorally, this diversity has been well characterized across many cultures and participants with data supporting the Moral Foundations Theory (Graham et al., 2009). Neurally, researchers have shown that different domains of morality are reflected by different neural activation (Parkinson et al., 2011). In this dissertation, we provide additional evidence for a diverse group of neural circuits, but with an enhanced approach for reverse inference. Such an approach allows not only the identification of separate neural elements, but also allows for the formal description of each circuit's function. We find that element 13 and element 24 are both important components of moral cognition, and we find associations with psychological processes of *mentalizing* and *taste*, respectively.

The stronger link between psychological process and neural circuit allows for the generation of hypotheses informed by behavior and neuroscience. For example, previous work has found that differing attitudes regarding domains of morality – such as harm, fairness, and purity – have been shown to be correlated with political affiliation (Graham et al., 2009; Haidt, 2007). Due to the associations between psychological process and our neural elements, we may hypothesize that the neural dynamics of element 13 (*mentalizing*) and element 24 (*taste*) may both predict political affiliation as well.

Even beyond these two primary neural circuits, we find that the diversity of morality is not limited only to mappings to various moral domains. Within the moral domain of harm, we highlight different mechanisms for blame and credit in Chapter 5. Emotion seems to play a larger role in blame and the ascription of intentionality, while statistical norms underlie credit and the denial of intentionality. These differences between negative and positive senses of morality have been corroborated by other studies (Takahashi et al., 2008), so even a further diversity and fragmentation of morality is evident if one permutes through each of the moral domains for combinations of negative and positive variants. To date, this is still an open project. All work done on the Knobe Effect has been limited to consequences involving physical or emotional harm. Future work may focus on whether the phenomenon holds true for purity violations, and whether there are distinct neural mechanisms for this domain.

Our work on the Knobe Effect also highlights a top-down role for neural processing in morality. Though it has been the common conception for intentionality to be an input into the higher-level process of moral judgment, it seems that there is a bottom-up and top-down loop. Such a dynamic has been shown in other domains of neuroscience, such as that involved in the top-down control of visual processing and memory by the prefrontal cortex (Zanto, Rubens, Thangavel, & Gazzaley, 2011). Such top-down influence can aid in narrowing or refining the range of computations that lower-level modules subsequently make, helping the system to accomplish a certain task with higher precision and efficiency.

In the realm of moral judgment, this type of modulation can also be very important. Presumably, the role of moral judgments and decisions helped our ancestors in cooperating with one another while avoiding others who could bring them harm. The fact that a moral judgment can influence assessments of intentionality can be viewed as a top-down adjustment of an intentionality threshold. The threshold for the judgment of an action as intentional is much lower for blame than it is for credit, and this asymmetry may ultimately reflect the fact that losses are often more important than gains in the goal of survival. Future work must be done to elucidate whether other components of morality such as emotion and value can also paradoxically be influenced by moral judgment in a top-down process. If so, these complex interactions between morality and its components make it harder to defend proposals of some single essence of morality.

### **6.3 Methodological Advancements**

We introduce two novel methodologies in the previous chapters. The first is the use of neuroscience to inform experimental philosophy. Experimental philosophy is itself a very new field, and it has fought hard to justify its use of empirical data for informing philosophical arguments (Knobe & Nichols, 2008; Nichols, 2011). The field's approach has been to survey what the *folk* have to say about philosophical issues that have mostly been debated throughout history by professional philosophers. By understanding what common intuitions are for these topics, experimental philosophers argue that they may identify ways in which the professional philosophers themselves may have been mistaken. For instance, Joshua Knobe takes this stance on the Knobe Effect, using his and others' empirical data on the matter to make a strong philosophical conclusion. He concludes that participants are not making any kind of error in their responses to the classic Knobe Effect scenarios, but rather, that the core competency that leads to intentionality judgments includes morality itself. This implies that the professional philosophers have unwittingly been making an error all along.

However, experimental philosophy's reliance on surveys of people's attitudes has some weaknesses. Some have pointed to the fact that such surveys are open to response biases and the inability of participants to accurately introspect (Carmel, 2011). In chapter 5, we attempted to address these concerns by bringing neuroscientific methods to bear on experimental philosophy. Just like the progression of behavioral

economics to neuroeconomics, we believe that such an approach can bring novel insights to both philosophy and neuroscience.

We also introduce a methodology for taking advantage of the rich dataset that is provided by Neurosynth. ICA composition has been previously done on a different meta-analytic database, *Brainmap*, but this database does not have the semantic precision of Neurosynth. Consequently, Smith et al. (2009) were able to make some broad claims about networks corresponding to emotion or attention. However, the semantic resolution could not allow for the distinction elements as specific as our *mentalizing* element or for the exploration of how certain elements build emergent compounds like *moral*.

#### **6.4 Practical Implications**

I will conclude with several proposals of practical implications for future work. First, the elements demonstrated in Chapter 3 can be used as an enhanced methodology for reverse inference. The reader may recall one loose end from Chapter 2, where the reverse inference (Table 3) for the DMPFC activation yielded a list of terms mostly consistent with *mentalizing*, but included some ambiguity within the analysis for the role of *default*. Instead of utilizing the decoder that is provided within the Neurosynth Core Tools Package (Yarkoni, 2013), we went back and performed a similar analysis using the 50 *neural elements*. The hypothesis was that these maximally independent spatial networks would represent an SPM of interest with higher specificity. This hypothesis

was supported: we found the distinction between the relevant reverse inferences to be higher using this analysis (Table 4). Element 13, the *mentalizing* component, is highly related to the activation from our study, and this association shows greater separation from the neural element representing resting state. This strengthens our claim that the DMPFC activation that tracked the magnitude of conflict of interest trial-by-trial in our task reflected our participants' engagement in mentalizing, presumably about the mental states of the patient. This also helps to strengthen the other conclusions from Chapter 2, which were based on this *mentalizing* reverse inference. The use of this methodology may be similarly useful in other cases in which researchers must interpret distributed networks of neural activity.

**Table 4: Decoding using the neural elements provides a higher degree of separation between competing reverse inferences in comparison to that shown in Table 3.**

<i>Neural Element</i>	Correlation
<b>DMPFC Tracking Patient Payments</b>	
<b>13 (<i>tom/mental/story/social/mentalizing</i>)</b>	<b>0.241</b>
26 ( <i>default/rest/restingstate</i> )	0.165
10 ( <i>reward</i> )	0.095
1 ( <i>motor/hand/movements/finger</i> )	0.085

The process of ROI selection could also be improved using the same logic. Traditionally, researchers have chosen an ROI from a previous study or set of studies, which relate to the psychological process of interest. However, this approach is subject to experimenter bias, since the choice of ROI is not purely based upon the nature of the psychological process, but also on some knowledge of the neural correlates of that

process. All of this could lead to circular reasoning. Are ROIs being chosen because of the associated psychological process or because a certain psychological process happens to activate the neural circuitry of interest?

Alternatively, researchers have also employed localizer tasks, which elicit certain neural circuitry by having participants directly engage in a psychological process. Activity observed within this localizer-defined ROI during a separate, unrelated task is then usually claimed as evidence for the presence of the psychological process that was relevant in the localizer task. However, the strength of the reverse inference in this case can also be tenuous (Poldrack, 2006). Assume that a localizer task activates a widely distributed neural network, such that many regions throughout the prefrontal cortex and the parietal cortices are activated. These areas of the brain are quite versatile, being active for a wide variety of psychological processes. Just because the localizer and the task of interest both elicited activation within the localizer region does not mean that the same underlying psychological process was associated with both.

Instead, the use of the 50 neural elements as ROIs affords two distinct advantages. First, the degrees of freedom for choosing a relevant ROI could be reduced compared to using maps directly from Neurosynth or from peak coordinates from other studies. For instance, let us say that one was interested in making a reverse inference about mentalizing in a certain study. The use of the neural elements provides the experimenter with one possible spatial map with which to proceed with the study. If the

experimenter were to use maps from Neurosynth, she would have the choice between multiple spatial maps associated with terms such as *tom*, *mental*, and *mentalizing*. Even worse, if the experimenter were to delve into the literature for reported coordinates from studies on the topic, the possible number of distinct ROIs that could be generated would expand even more substantially. Ultimately, the reduction in the degrees of freedom in ROI selection could aid in reducing false positives.

Additionally, the one neural network for mentalizing was formed based on reverse inference data from the Neurosynth database, allowing for stronger claims of specificity. More promiscuous regions, which activate for many psychological processes across the literature, do not show up in reverse inference maps because of the Bayesian posterior probability that is employed in the formation of these maps (Yarkoni et al., 2011). Accordingly, the peaks of the spatial map corresponding to the mentalizing element represent the coordinates that have the highest specificity for mentalizing as opposed to any other term in the Neurosynth database. One related aspect to this specificity is the fact that the 50 neural elements are maximally spatially independent, allowing for stronger distinctions between elements, as demonstrated by comparing Table 3 and Table 4.

Finally, I will end with distinct implications of the previous work for the application to medicine, particularly the approach to mental health and disorders. Within psychiatric research, there has been increasing interest in moving away from

using clinical definitions of mental disorders as defined by the Diagnostic and Statistical Manual (DSM) towards the use of the analysis of distinct domains. Such domains have been proposed by the National Institute for Mental Health's (NIMH) Research Domain Criteria (RDoC), and include cognition, emotion, and behavior (NIMH, 2011). In the context of this change, two future avenues of research would be relevant.

The RDoC proposes various units levels of analysis for approaching different domains and constructs, ranging from genes to behaviors. One such level between these two extremes is the neural circuit, and their study could be enhanced with the use of the neural elements summarized in Table 5. For instance, the use of element 13 (*mentalizing*) could prove to be useful for researchers studying autism spectrum disorder (ASD). ASD has long been theorized as a disorder arising from deficits of theory-of-mind (Baron-Cohen, Leslie, & Frith, 1985; Frith, 2001). This has been reflected neurally: studies with Asperger's syndrome participants have shown deficits within the mentalizing network compared to controls (Baron-Cohen et al., 1999). Though such a perspective has been the most common approach to autism, other studies have also surprisingly found that there may be deficits in activity in reward networks (Zeeland & Ashley, 2010). Such reward networks are strongly associated with the brain regions represented by element 10 (*reward*). As researchers continue to dissect the mechanisms of disorders like ASD, they may find benefit in using these *neural elements* in their analyses of mentalizing and

reward circuitries. Additionally, the *neural elements* could guide the exploration of other circuits not yet associated with a disease.

The second major theme of this dissertation points to the types of tasks that could be used to elicit activity in these neural elements. For many mental disorders, the use of tasks involving moral judgments and decisions may prove to be an extremely effective. Several studies have analyzed moral judgment from various neural components in isolation. Compared to control participants, Asperger's syndrome participants have been shown to rely less on mentalizing in the formation of moral judgments (Moran et al., 2011). It seems that the neural representations of such computations in patient populations are also distinct (Koster-Hale et al., 2013). In a different domain, emotional processing has been shown to be dysfunctional in patients with psychopathy (Blair, 2007).

Moving forward, future studies could leverage the core advantage of using moral paradigms: its ability to elicit the activity of *multiple* neural elements, which interact with one another to support an emergent neural computation. Instead of looking at the role of each of these neural elements in morality in isolation, it may be that the interaction between these elements could bring novel insights into a wide range of mental disorders. Indeed, morality may hold a distinct advantage as a target of study compared to many other possible *neural compounds*. Morality seems to elicit a remarkable number of domains cited by the NIMH and summarized in Table 5.

**Table 5: Proposed mapping between domains proposed by the NIMH's RDoC and the neural elements described in Chapter 3.**

<i>NIMH RDoC Domain/Construct</i>	<i>Relevant Neural Elements</i>
Negative Valence Systems	3 ( <i>emotion/neutral</i> ) 32 ( <i>negative/positive</i> )
Positive Valence Systems	10 ( <i>reward</i> ) 51 ( <i>positive/negative</i> )
Visual Perception	9 ( <i>visual</i> ) 20: ( <i>motion/visual</i> ) 48 ( <i>audiovisual/integration</i> )
Auditory Perception	1 ( <i>auditory</i> ) 48 ( <i>audiovisual/integration</i> )
Olfactory/Somatosensory/Multimodal	24 ( <i>taste/rating/food/eating/olfactory/reward/physiological</i> )
Declarative Memory	21 ( <i>remembered/encoded/remember</i> ) 23 ( <i>recollection</i> )
Language Behavior	4 ( <i>phonology/word</i> ) 6 ( <i>sentence</i> ) 35 ( <i>readers/reading</i> ) 40 ( <i>semantic</i> ) 55 ( <i>verbal/verb</i> ) 62 ( <i>verb/verbal</i> )
Working Memory	16 ( <i>work/load/working/2back/memory/execute/1back</i> )
Facial Communication	42 ( <i>face/facial/social/expression</i> )
Understanding Mental States	13 ( <i>tom/mental/story/social/mentalizing</i> )

There may be a core reason for this. For millennia, morality has guided the survival of humans and the course of progress of societies. For some part of this history, academics have similarly been fascinated by its nature and its essence. Today, even a brief look at the front pages at any newsstand captures our attention because its headlines almost all have a fundamentally moral nature. Is it wrong for country X to

invade country Y? Is it wrong that celebrity X cheated on celebrity Y? A recent paper has provided several converging lines of evidence suggesting that moral traits form the most central and important part of what people view to be one's identity. More so than memories, desires, perceptions, and physical traits, our personal identity is most essentially composed of our moral character (Strohminger & Nichols, 2014). It is not surprising, then, that a promising way to proceed in better understanding who we are, whether in health or illness, is to delve more deeply into who we are as moral creatures.

## References

- Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind? *NeuroImage*, *30*(3), 1059–68.  
doi:10.1016/j.neuroimage.2005.10.026
- Alexander, J. (2012). *Experimental Philosophy: An Introduction*. *Philosophia* (Vol. 40, p. 200). Polity. doi:10.1007/s11406-012-9392-3
- Alicke, M. (2008). Blaming Badly. *Journal of Cognition and Culture*, *8*(1-2), 179–86.
- Aquinas, T. (1988). Summa Theologica. In R. Regan & W. Baumgarth (Eds.), *On Law, Morality and Politics* (pp. 226–227). Indianapolis/Cambridge: Hackett Publishing Co.
- Aristotle. (1959). *The Nichomachean Ethics*. (R. WD, Ed.). London: Oxford University Press.
- Augustine, S. B. of H. (1947). *De libero arbitrio voluntatis St. Augustine on free will*. Charlottesville: University of Virginia.
- Aziz-Zadeh, L., Kaplan, J. T., & Iacoboni, M. (2009). “Aha!”: The neural correlates of verbal insight solutions. *Human Brain Mapping*, *30*(3), 908–16.  
doi:10.1002/hbm.20554
- Balleine, B. W., & O’Doherty, J. P. (2010). Human and rodent homologues in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69. doi:10.1038/npp.2009.131
- Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*.
- Baron-Cohen, S., Ring, H. a, Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, a, & Williams, S. C. (1999). Social intelligence in the normal and autistic brain: an fMRI study. *The European Journal of Neuroscience*, *11*(6), 1891–8.
- Baumeister, R. F., & Vonasch, A. J. (2012). Is the Essence of Morality Mind Perception, Self-Regulation, Free Will, or Culture? *Psychological Inquiry*, *23*(2), 134–136.  
doi:10.1080/1047840X.2012.667758

- Beckmann, C. F., & Smith, S. M. (2004). Functional Magnetic Resonance Imaging, 23(2), 137–152.
- Bentham, J. (1907). *An Introduction Principles of Morals and Legislation*. Oxford: Clarendon Press.
- Berker, S. (2009). The Normative Insignificance of Neuroscience. *Philosophy & Public Affairs*, 37(4), 293–329. doi:10.1111/j.1088-4963.2009.01164.x
- Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition*, 57(1), 1–29. doi:10.1016/0010-0277(95)00676-P
- Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, 11(9), 387–92. doi:10.1016/j.tics.2007.07.003
- Brown, R. (2003). Measuring Individual Differences in the Tendency to Forgive: Construct Validity and Links With Depression. *Personality and Social Psychology Bulletin*, 29(6), 759–71.
- Buckholz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, 60(5), 930–40. doi:10.1016/j.neuron.2008.10.016
- 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. doi:10.1177/1745691610393980
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550–558. doi:10.1037/a0018933
- Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function*, 217(4), 783–96. doi:10.1007/s00429-012-0380-y
- Cain, D. M., Loewenstein, G., & Moore, D. A. (2005). The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest. *The Journal of Legal Studies*, 34(1), 1–25. doi:10.1086/426699

- Cain, D. M., Loewenstein, G., & Moore, D. A. (2011). When Sunlight Fails to Disinfect: Understanding the Perverse Effects of Disclosing Conflicts of Interest. *The Journal of Consumer Research*, 37(5), 836–857. doi:10.1086/656252
- Camerer, C. (2003). *Behavioral game theory: experiments in strategic interaction*. New York [u.a.]: Russell Sage [u.a.].
- Carmel, D. (2011). Experimental philosophy: surveys alone won't fly. *Science*, 332(6035), 1262; author reply 1262–3. doi:10.1126/science.332.6035.1262-b
- Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, 337(6090), 109–111. doi:10.1126/science.1219681
- Carver, C., & White, T. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending rewards and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, 67, 319–33.
- Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 5(5), 411–419.
- Christie, R., & Geis, F. (1970). *Studies in Machiavellianism*. New York: Academic Press.
- Cokely, E., & Feltz, A. (2009). Individual differences, judgment biases, and theory-of-mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality*, 43(1), 18–24. doi:10.1016/j.jrp.2008.10.007
- Costa, P., & MacCrae, R. (1992). Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual. *Psychological Assessment Resources*.
- Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, 17(8), 363–6. doi:10.1016/j.tics.2013.06.005
- Cushman, F., & Mele, A. (2008). Intentional Action: Two-and-a-half Folk Concepts? In *Experimental Philosophy* (pp. 171–188). New York: Oxford University Press.
- Davis, M. M. H., & Association, A. P. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10(4), 85.

- Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, 8(11), 1611–8. doi:10.1038/nn1575
- Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 24(8), 1742–52. doi:10.1162/jocn\_a\_00233
- Falk, A., & Szech, N. (2013). Morals and markets. *Science*, 340(6133), 707–11. doi:10.1126/science.1231566
- Feltz, A. (2007). The Knobe Effect: A Brief Overview. *Journal of Mind and Behavior*, 28, 265–77.
- FMRIB. (2014). FMRIB Software Library.
- Foot, P. (1978). *Virtues and vices and other essays in moral philosophy*. Berkeley: University of California Press.
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. doi:10.1257/089533005775196732
- Frith, U. (2001). Mind blindness and the brain in autism. *Neuron*, 32, 969–979.
- Graham, J., Haidt, J., & Nosek, B. a. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029–46. doi:10.1037/a0015141
- Grant, C. M., Boucher, J., Riggs, K. J., & Grayson, A. (2005). Moral understanding in children with autism. *Autism: The International Journal of Research and Practice*, 9(3), 317–31. doi:10.1177/1362361305055418
- Gray, K., Young, L. L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, 23(2), 101–124. doi:10.1080/1047840X.2012.651387
- Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322–3; author reply 323–4. doi:10.1016/j.tics.2007.06.004

- Greene, J. D. (2010). The Secret Joke of Kant's Soul. *Moral Psychology: Historical and Contemporary Readings*, 359.
- Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517–523.
- Greene, J. D., Morelli, S. a, Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144–54. doi:10.1016/j.cognition.2007.11.004
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400. doi:10.1016/j.neuron.2004.09.027
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–8. doi:10.1126/science.1062872
- Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality & Social Psychology Bulletin*, 36(12), 1635–47. doi:10.1177/0146167210386733
- Haidt, J. (2001). The Emotional Dog and Its Rational Tail □: A Social Intuitionist Approach to Moral Judgment, *108(4)*, 814–834. doi:10.1037//0033-295X.
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 998(2007). doi:10.1126/science.1137651
- Haidt, J., & Joseph, C. (2004). Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4), 55–66.
- Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169), 557–9. doi:10.1038/nature06288
- Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, 316(5831), 1622–5. doi:10.1126/science.1140738
- Hare, R. (2003). *The Hare psychopathy checklist-revised*. North Tonawanda N.Y.: Multi-Health Systems Inc.

- Hart, H. (1968). *Punishment and responsibility: essays in the philosophy of law*. New York: Oxford University Press.
- Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *Science*, 320(5879), 1092–5. doi:10.1126/science.1153651
- Hume, D. (1978). *A treatise of human nature*. Cambridge [Cambridgeshire]; New York: Clarendon Press;Oxford University Press.
- Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3), 626–34. doi:10.1109/72.761722
- Blair, R. J. R. (1996). Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders*, 26(5), 571–579. doi:10.1007/BF02172277
- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2), 825–41. doi:10.1006/nimg.2002.1132
- Kant, I. (2002). *Groundwork for the Metaphysics of Morals*. (A. Wood, Ed.) (p. 194). Yale University Press.
- Kelley, K. (n.d.). MBESS: An R Package. Retrieved from <http://www3.nd.edu/~kkelley/site/MBESS.html>
- Kelly, M., Ngo, L., Huettel, S. A., & Sinnott-Armstrong, W. (2014). *Social Influence in Online Interaction*.
- 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(9), 677–684. doi:10.1016/S0006-3223(01)01222-7
- Kliemann, D., Young, L. L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, 46(12), 2949–57. doi:10.1016/j.neuropsychologia.2008.06.010
- Knobe, J. (2003). Intentional Action and Side Effects of Ordinary Language. *Analysis*, 63(3), 190–94. doi:10.2307/3328244

- Knobe, J. (2005). Theory of mind and moral cognition: exploring the connections. *Trends in Cognitive Sciences*, 9(8), 357–9. doi:10.1016/j.tics.2005.06.011
- Knobe, J. (2010). Person as scientist, person as moralist. *The Behavioral and Brain Sciences*, 33(4), 315–29; discussion 329–65. doi:10.1017/S0140525X10000907
- Knobe, J., & Burra, A. (2006). The Folk Concepts of Intention and Intentional Action: A Cross-Cultural Study. *Journal of Cognition and Culture*, 6(1), 113–32. doi:10.1163/156853706776931222
- Knobe, J., & Nichols, S. (2008). *Experimental Philosophy*. Oxford University Press.
- Koenigs, M., Young, L. L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908–11. doi:10.1038/nature05631
- Kohlberg, L. (1984). *The psychology of moral development: the nature and validity of moral stages*. San Francisco: Harper & Row.
- Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, 110(14), 5648–53. doi:10.1073/pnas.1207992110
- Larsen, R. J. (1984). Theory and measurement of affect intensity as an individual difference characteristic. *Dissertation Abstracts International*, 85(2297B).
- Leben, D. (2010). Cognitive Neuroscience and Moral Decision-making: Guide or Set Aside? *Neuroethics*, 4(2), 163–174. doi:10.1007/s12152-010-9087-z
- Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, 1–12. doi:10.1016/j.conb.2012.06.001
- Loewenstein, G., Sah, S., & Cain, D. M. (2012). The unintended consequences of conflict of interest disclosure. *JAMA*, 307(7), 669–70. doi:10.1001/jama.2012.154
- Lohr, D. (n.d.). Pedro Lopez: The Monster of the Andes. *Crime Library*.
- Machery, E. (2008). The Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind & Language*, 23(2), 165–189. doi:10.1111/j.1468-0017.2007.00336.x

- Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: the tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences & the Law*, 21(5), 563–80. doi:10.1002/bsl.554
- Mallon, R. (2008). Knobe vs. Machery: Testing the Trade-Off Hypothesis. *Mind and Language*, 23(2), 247–55.
- McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). Profile of Mood States. *San Diego, California: Educational and Industrial Testing Service*.
- Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, 18(4), 193–7.
- Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, 11(4), 143–52. doi:10.1016/j.tics.2006.12.007
- Mill, J. (2002). *The Basic Writings of John Stuart Mill*. New York: Modern Library.
- Miller, M. B., Sinnott-Armstrong, W., Young, L. L., King, D., Paggi, A., Fabri, M., ... Gazzaniga, M. S. (2010). Abnormal moral reasoning in complete and partial callosotomy patients. *Neuropsychologia*, 48(7), 2215–20. doi:10.1016/j.neuropsychologia.2010.02.021
- Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, 18(2), 262–71. doi:10.1093/cercor/bhm051
- Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional Networks in Emotional Moral and Nonmoral Social Judgments. *NeuroImage*, 16(3), 696–703. doi:10.1006/nimg.2002.1118
- Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience*, 22(7), 2730–6. doi:20026214
- Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignácio, F. A., Bramati, I. E., Caparelli-Dáquer, E. M., & Eslinger, P. J. (2005). The moral affiliations of disgust: a functional MRI study. *Cognitive and Behavioral Neurology*, 18(1), 68–78.

- Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*(1), 33–43.
- Monroe, A. E., Guglielmo, S., & Malle, B. F. (2012). Morality Goes Beyond Mind Perception. *Psychological Inquiry, 23*(2), 179–184. doi:10.1080/1047840X.2012.668271
- Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O’Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences, 108*(7), 2688–92. doi:10.1073/pnas.1011734108
- Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron, 75*(1), 73–9. doi:10.1016/j.neuron.2012.05.021
- Murder in the first degree, N.Y State Penal Law Section § 125.27.
- Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations, 9*(2), 203–219. doi:10.1080/13869790600641905
- Newton, M. (2000). *The Encyclopedia of Serial Killers*. New York: Facts on File.
- Nichols, S. (2011). Experimental Philosophy and the Problem of Free Will. *Science, 331*(March), 1401–1403. doi:10.1126/science.1192931
- Nichols, S., & Ulatowski, J. (2007). Intuitions and Individual Differences: The Knobe Effect Revisited. *Mind & Language, 22*(4), 346–365. doi:10.1111/j.1468-0017.2007.00312.x
- NIMH. (2011). NIMH Research Domain Criteria (RDoC). Retrieved from <http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml>
- Norman, R. (1995). *Ethics, killing, and war*. Cambridge; New York N.Y.: Cambridge University Press.
- Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*(6), 972–87.

- Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–19.
- Parker, I. (2004, August 2). The Gift. *New Yorker*, pp. 54–63.
- Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E. P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience*, 23(10), 3162–80. doi:10.1162/jocn\_a\_00017
- Phelan, M., & Sarkissian, H. (2008). The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138(2), 291–98. doi:10.1007/s11098-006-9047-y
- Piaget, J. (1977). *The essential Piaget*. New York: Basic Books.
- Pinillos, N. Á., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy's New Challenge: Experiments and Intentional Action. *Mind & Language*, 26(1), 115–139.
- Platt, M. L., & Huettel, S. A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nature Neuroscience*, 11(4), 398–403. doi:10.1038/nn2062
- Poldrack, R. a. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2), 59–63. doi:10.1016/j.tics.2005.12.004
- Posner, R. (1992). *Sex and reason*. Cambridge: Cambridge University Press.
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36(4), 717–31.
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, 16(2), 93–115. doi:10.1037/a0022658
- Premack, D., & Premack, a J. (1997). Infants Attribute Value± to the Goal-Directed Actions of Self-propelled Objects. *Journal of Cognitive Neuroscience*, 9(6), 848–56. doi:10.1162/jocn.1997.9.6.848
- Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, 25(1), 1–20; discussion 20–71.

- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1).
- Rand, D. G. (2012). The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299, 172–9. doi:10.1016/j.jtbi.2011.03.004
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427–30. doi:10.1038/nature11467
- Sanbar, S. (2004). *Legal medicine*. St. Louis: Mosby.
- SAS. (2011). *SAS/STAT 9.3 User's Guide*.
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–9. doi:10.1111/j.1467-9280.2006.01768.x
- Saxe, R., & Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43(10), 1391–9. doi:10.1016/j.neuropsychologia.2005.02.013
- Schaich Borg, J., Hynes, C., Horn, J. Van, Grafton, S., Sinnott-Armstrong, W., & Van Horn, J. (2006). Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation. *Journal of Cognitive Neuroscience*, 18(5), 803–17. doi:10.1162/jocn.2006.18.5.803
- Schaich Borg, J., Lieberman, D., & Kiehl, K. a. (2008). Infection, incest, and iniquity: investigating the neural correlates of disgust and morality. *Journal of Cognitive Neuroscience*, 20(9), 1529–46. doi:10.1162/jocn.2008.20109
- Schluchter, M. D., & Elashoff, J. T. (1990). Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computation and Simulation*, 37(1-2), 69–87.
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality & Social Psychology Bulletin*, 34(8), 1096–109. doi:10.1177/0146167208317771
- Scipy. (2013).

- Shapiro, D. N., Chandler, J., & Mueller, P. a. (2013). Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science, 1*(2), 213–20. doi:10.1177/2167702612469015
- Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–77. doi:10.1016/j.neuron.2010.07.020
- 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), 1157–62. doi:10.1126/science.1093535
- Sinnott-Armstrong, W. (2012). Does Morality Have an Essence? *Psychological Inquiry, 23*(2), 194–197. doi:10.1080/1047840X.2012.666653
- Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping, 17*(3), 143–55. doi:10.1002/hbm.10062
- Smith, S. M., Fox, P. T. M., Miller, K. L., Glahn, D. C., Mackay, C. E., Filippini, N., ... Beckmann, C. F. (2009). Correspondence of the brain’s functional architecture during activation and rest. *Proceedings of the National Academy of Sciences, 106*(31), 13040–5. doi:10.1073/pnas.0905267106
- Snijders, T. A., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: SAGE Publications Inc.
- Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489–510. doi:10.1162/jocn.2008.21029
- Sripada, C. S. (2009). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies, 151*(2), 159–76. doi:10.1007/s11098-009-9423-5
- Strohming, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition, 119*(2), 295–300. doi:10.1016/j.cognition.2010.12.012
- Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131*(1), 159–71. doi:10.1016/j.cognition.2013.12.005

- Stuckless, N., & Goranson, R. (1992). The Vengeance Scale: Development of a measure of attitudes toward revenge. *Journal of Social Behavior & Personality*, 7(1), 25–42.
- Stump, E. (2001). *The Cambridge Companion to Augustine*. Cambridge UK; New York: Cambridge University Press.
- Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4), 814–20.  
doi:10.1016/j.shpsc.2012.05.009
- Takahashi, H., Kato, M., Matsuura, M., Koeda, M., Yahata, N., Suhara, T., & Okubo, Y. (2008). Neural correlates of human virtue judgment. *Cerebral Cortex*, 18(8), 1886–91.  
doi:10.1093/cercor/bhm214
- Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences*, 109(21), 8038–43.  
doi:10.1073/pnas.1202129109
- Tankersley, D., Stowe, C. J., & Huettel, S. A. (2007). Altruism is associated with an increased neural response to agency. *Nature Neuroscience*, 10(2), 150–1.  
doi:10.1038/nn1833
- Thomson, J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist*, 59(2), 204–217.
- Tsukiura, T., & Cabeza, R. (2011). Shared brain activity for aesthetic and moral judgments: implications for the Beauty-is-Good stereotype. *Social Cognitive and Affective Neuroscience*, 6(1), 138–48. doi:10.1093/scan/nsq025
- Turiel, E. (1983). *The development of social knowledge: morality and convention*. Cambridge [Cambridgeshire]; New York: Cambridge University Press.
- Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1), 87–100.  
doi:10.1016/j.cognition.2010.04.003
- Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, 17(6), 476–7. doi:10.1111/j.1467-9280.2006.01731.x
- White, J. (1985). *Contemporary moral problems*. St. Paul: West Pub. Co.

- Wittmann, M., Simmons, A. N., Aron, J. L., & Paulus, M. P. (2010). Accumulation of neural activity in the posterior insula encodes the passage of time. *Neuropsychologia*, 48(10), 3110–20. doi:10.1016/j.neuropsychologia.2010.06.023
- Woodward, P. A. (2001). *The Doctrine of Double Effect: Philosophers Debate a Controversial Moral Principle* (p. 317). University of Notre Dame Press.
- Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, 14(6), 1370–86. doi:10.1006/nimg.2001.0931
- Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., & Evans, a C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15(1), 1–15. doi:10.1006/nimg.2001.0933
- Wright, J. C., & Bengson, J. (2009). Asymmetries in Judgments of Responsibility and Intentional Action. *Mind & Language*, 24(2004), 24–50.
- Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, 15(5), 786–91. doi:10.1038/nn.3068
- Yarkoni, T. (2011). Neurosynth. Retrieved February 04, 2013, from <http://www.neurosynth.org>
- Yarkoni, T. (2013). Neurosynth Repository on Github. Retrieved from <https://github.com/neurosynth/neurosynth>
- Yarkoni, T., Poldrack, R., & Nichols, T. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8, 665–74.
- Young, L. L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15), 6753–8. doi:10.1073/pnas.0914826107
- Young, L. L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, 6(1-2), 291–304.

- Young, L. L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, 104(20), 8235–40. doi:10.1073/pnas.0701408104
- Young, L. L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, 21(7), 1396–405. doi:10.1162/jocn.2009.21137
- Zanto, T. P., Rubens, M. T., Thangavel, A., & Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience*, 14(5), 656–61. doi:10.1038/nn.2773
- Zeeland, S., & Ashley, A. (2010). Reward processing in autism. *Autism Research*, 3(2), 53–67. doi:10.1002/aur.122.Reward

## **Biography**

Lawrence Ngo was born in Fountain Valley, California in 1986. He double majored in Philosophy and Biology at Wake Forest University, graduating Phi Beta Kappa and summa cum laude in 2008. He was supported by the Kenan Institute for Ethics Graduate Fellowship and the Wakeman Fellowship in Neurobiology. He will complete one remaining year of medical training in the School of Medicine before doing an internship in Internal Medicine and residency in Radiology with an intended fellowship in Neuroradiology.