How Many Moralities? a Bottom-Up Approach to Mapping the Brain's Natural Moral Categories

How Many Moralities?

A Bottom-Up Approach to Mapping the Brain's Natural Moral Categories

Michael T. Gravina

A Thesis in the Field of Biology

for the Degree of Master of Liberal Arts in the Extension Studies

Harvard University

November 2015

Abstract

The external structure and internal boundaries of the moral domain are not sharply defined. Substantive definitions of morality struggle to cleanly encapsulate the full diversity of human moral concern without including too much to retain correspondence to folk understandings, while functionalist definitions are complex and difficult to implement in study. Psychological work in 20th century often assumed morality was a single domain concerned primarily with transgression types emphasized in Western academia. Recent brain-imaging work has suggested that morality may in fact comprise multiple sub-domains, corresponding to moral natural kinds which cover a more diverse spectrum of topics than Western morality is typically concerned with (Parkinson et al., 2011). Moral Foundations Theory (MFT), which takes an evolutionary functionalist approach, is a promising candidate structure for this expanded moral domain.

Here I probe the structure of the moral domain in exploratory fashion for correlates to the foundations of MFT in patterns of brain activation in response to moral stimuli generated and categorized by survey respondents. Activation contrasts are used to identify regions of differential activity between the putative foundations. Conjunctive overlaps between foundation contrasts are compared in order to establish which foundations behave similarly to one another relative to the other foundations in the set.

Neither the 5-factor structure of MFT nor its coarser 2-factor structure is upheld. Instead, a semi-polarized scheme is suggested, with harm-preventative and purity-maintaining moral types occupying the extremes and more interpersonal foundations grouped together in between and less clearly delineated than previously assumed.

Dedication

Dedicated to Genie Giaimo, who supports me without reservation, who inspires me by example, and who, by believing in me so intensely, has helped me to believe in myself.

Acknowledgments

I would like to thankfully acknowledge Alek Chakroff, whose role in the genesis of this project and whose guidance during its execution cannot be understated. This work is best considered a narrow elaboration on fine points, the broader context of which Alek investigated in much more exhaustive fashion in his dissertation – so thank you, Alek, for providing the giant shoulders for me to stand on. Also, to Dr. Greene, genuine gratitude both for granting me this opportunity in the first place and for shepherding me through to its conclusion.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

If you were to ask any person off the street whether or not it was moral to kill someone, or to steal, disrespect one's elders, commit treason or blaspheme, chances are they would produce an answer quickly. Ask them instead not whether a particular action is moral, but what topics can be talked about in moral terms in the first place, and they might take slightly longer to answer, but they would certainly be able to come up with a list, and that list would have considerable overlap with those offered by your other respondents. It would be very surprising to find, for instance, someone who considered the tying of shoes to be either righteous or transgressive while thinking that assault with a deadly weapon is morally irrelevant. There clearly are boundaries to the behaviors which humans judge as either moral or immoral, but that does not mean that those boundaries are themselves clear; formally speaking, the internal and external dividing lines of morality are still poorly defined.

Even the basics are hotly contested. Morality is considered to apply to *actions* in deontological and consequentialist constructions, but virtue-based ethical philosophies are instead primarily concerned with characteristics of the *agent* who performs the action (Hursthouse 2012). It tends to be concerned with *interactions* between people or groups, and, according to seminal studies by evolutionary psychologists such as Cosmides & Tooby, tends to involve a *harm* of some sort (Cosmides & Tooby, 2005; Tooby & Cosmides, 2008; Tooby & Cosmides, 2010) - yet cross-cultural studies have documented a significant subset of moral beliefs, concerned with matters of physical purity and spiritual sanctity, which are not necessarily either inter-personal or related to any

concrete harm (Haidt 2012). Jonathan Haidt defines morality by function rather than subject: "Moral systems are interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate self-interest and make cooperative societies possible" (Haidt 2012). This functionalist definition offers to reconcile divisions between contending camps in moral theory but is complex, context-bound, and difficult to prospectively outline.

Such theoretical complexities, combined with improvements in brain-imaging technology, have motivated the search for more directly observable structure in the moral domain as implemented in the brain. Early work in cognitive neuroscience attempted to identify domain-specific brain networks for morality in the same way that other specialized regions might process the domain of faces or the domain of language. But recent study suggests that, instead of a single, homogeneous "moral domain," there may be a more complex cognitive organizational structure for morality which comprises *multiple* domains (Shweder, 1997; Parkinson et al., 2011; Haidt, 2012). From the perspective of the brain, then, there may be such a thing as distinct moral "natural kinds."

Existing paradigms in moral domain modeling

Early models had few divisions and were oriented towards harm-prevention and justice

Mainstream treatments of morality as a more or less unified domain of human concern can be traced back at least to the writings of 19th century utilitarian philosophers like John Stuart Mill, who postulated that all moral particulars were derived from a very

small number of principles, particularly the prevention of harms (Haidt 2012). Though Mill's was not a description of human psychological faculties *per se*, but instead of logical justifications for moral principles, it set a unitary tone for the study of morality which would strongly color later explorations of the moral mind. In the 20th century, psychologists Lawrence Kohlberg and Eliot Turiel codified the view that harm-prevention and justice (the latter concept synthesizing the related principles of fairness and individual rights) are fundamental moral rules, that they are universal targets of high-level moral development in humans, and that all other moral rules are to be derived by rational extrapolation from these deontic starting points. Any rules which can not be reduced to harm-prevention and justice, including prohibitions on "victimless acts" like sexual deviance or defacement of a national flag, are, in this framework, merely social conventions rather than genuine morality. In the 1990s, psychologists and cultural anthropologists, using behavioral methods to search for universals in the human moral repertoire, confirmed that certain moral transgression types, comprising physical harms and the misappropriation of resources, were indeed condemned across all cultures. (Shweder, Much, Mahapatra & Park, 1997). These types of transgressive behavior, linked in that they both have negative impacts on direct interpersonal cooperation, correspond strongly with the harm-preventative and justice-based formulations of the tradition.

When brain-imaging technology came of age in the late 1990s and early 2000s, psychology's focus on one- or two-factor paradigms implicitly shaped the thrust of research into the functional neuroscience of morality. Theoretical and empirical research accumulated on the evolutionary forces (Axelrod, 2012) and neural mechanisms

underlying the harm and justice aspects of morality. Research during this period established a basic human aversion to causing physical harm (Cushman, Gray, Gaffey & Mendes, 2012), a visceral disgust reaction to violations of fairness and equity (Chapman, Kim, Susskind, & Anderson, 2009; Cannon, Schnall, & White, 2011) and motivations towards both positive and negative forms of altruism, or the assumption of personal costs to help sympathetic others and impose sanctions on norm-violators (De Quervain et al., 2004; Hamann, Warneken, Greenberg, & Tomasello, 2011). In general, the field was dominated by inquiries into the harm and justice aspects of morality, plausibly reflecting the influence of an antiquated and excessively restricted construction of the moral domain (Graham et al., 2011).

Challenges to WEIRD models

Models of the moral domain such as these extrapolate the moral relevance of actions from their consequences, framed in terms of injuries or gains by individuals, rather than by reference to larger social groups or human-transcendent anchors of value. They have enjoyed an easy ideological fit with the cultures of their proponents, which are predominantly "WEIRD" - Western, Educated, Industrial, Rich, and Democratic. But the WEIRD culture which envelops the scholastic communities most active in moral neuropsychology is far from the norm in global human communities. By reifying normative variations as moral or instead consigning them to merely conventional status based chiefly on their endorsement by Western academic tradition, these models may define away the true scope of morality in the majority of the world and even among the more politically and religiously conservative enclaves of the West itself (Haidt 2012).

But if substantive definitions of morality are to be rejected to help protect against the risk of cultural chauvinism, they must be replaced with some other system which can be applied independent of culture. One such approach is to characterize the boundaries of moral judgments and their accompanying sanctioning behaviors by their shared *functional* nature, regulating the selfishness of individuals in order to make life within social super-organisms possible. According to the functionalist paradigm, moralities co-evolve with culture to suit the demands of the local human environment. We should, in this view, not expect to find the same assortment or emphasis of moral types across all cultures as we find in liberal academic settings in the West (Haidt, 2007; Haidt & Kesebir, 2010). The dawning recognition of WEIRD bias called a new wave of researchers to reformulate the moral domain in a more inclusive form based on such functional considerations.

Simultaneously, a related assumption in neuroscience, that the neural correlates of moral cognition are well-integrated enough to be meaningfully studied as a unitary function of the brain, was coming under scrutiny as a number of theorists focused on divisions and structure within moral judgments (Parkinson et al., 2011). By characterizing underlying structural constraints and dimensions, such as the judged severity of violations, (Cushman, 2006), these researchers began to characterize the interpretation of moral propositions as a series of connected processing steps and, even in cases where the researchers themselves argued for unified morality, such as Mikhail's theory of a universal computational grammar of morally relevant statements (2007), still paved the way for more reductive, taxonomic schemes for classifying the variation in moral processing.

Richard Shweder, working with his collaborators Much, Mahapatra, and Park, was one of those aiming to integrate a fuller range of moral concerns than the WEIRD tradition allowed. Interviewing the mostly Brahmin Hindu residents of Orissa, India about moral violations, he analyzed the interview transcripts through "inductive and iterative reading… and classification" in order to identify consistent moral themes. While Shweder did observe prominent concern for harm and fairness in common with WEIRD subjects, he also found markedly greater moral concern on subjects relating to communal cohesiveness and spiritual purity. Based on cluster analysis of the classifications, Shweder proposed an expansion of the moral domain into three "ethics": Autonomy (concerned with interactions between individuals and corresponding to harm-prevention and fairness), Community (concerned with the social hierarchy and the obligations of individuals to the greater whole), and Divinity (concerned with maintenance of purity and spiritual elevation) (Shweder et al., 1997). Empirical support in line with Shweder's framework exists but is sparse: evidence linking each of the three proposed ethics with a corresponding, socially-oriented emotion (anger with violations of Autonomy, contempt with violations of Community, and disgust with violations of Divinity) provides a plausible mechanism for their behavioral implementation, (Rozin, Lowery, Imada & Haidt, 1999) while findings of distinct neural correlates for transgressions of each ethic (though not explicitly for the ethics themselves, as positively stated) argue against simpler functional taxonomies, including the concept of a unified "morality module." (Parkinson et al., 2011).

A successful model of moral cognition must, in addition to comprehensively parsing the moral domain, also make a compelling account of the evolutionary forces

which defined the putative areas of concern. Starting from Shweder's framework, Jonathan Haidt and Craig Joseph used a "meta-empirical method" to explore the topic, surveying theoretical literature for moral concerns which transcend cultural boundaries and then splitting these concerns into categories. However, Haidt and Joseph added an additional criterion that any supposed universal moral category have a clear counterpart in the adaptive challenges predicted by evolutionary psychology. For example, the large body of work on reciprocal altruism meant that fairness of interpersonal transactions would almost certainly comprise a moral type.

The result of this work was an expansion from Shweder's 3 ethics to a new, 5-factor Moral Foundations Theory (MFT). The foundations of MFT are mostly finer divisions of Shweder's ethics: the ethic of Autonomy is divided into the foundation of Harm/Care (based on the adaptive challenge of caring for children) and the foundation of Fairness/Reciprocity (based on the challenge of benefiting from cooperation with conspecifics without being exploited), the ethic of Community becomes the foundations of Ingroup/Loyalty (building and sustaining coalitions) and Authority/Respect (forming beneficial relationships within social hierarchies), and the ethic of Divinity is translated almost wholesale into the foundation of Purity/Sanctity (based on the communal challenge of remaining free of noxious contaminants and generalized to include "contaminating" behaviors and symbols) (Haidt, 2012).[1] With Jesse Graham, Haidt developed the Moral Foundations Questionnaire to assay endorsement of these 5

[1] For the purpose of succinctness, I refer to the 5 foundations of MFT from here on as Care, Fairness, Loyalty, Respect, and Sanctity, except when discussing MFT as explored by its authors, in which case I use the original foundation labels.

foundations in large, heterogeneous populations, and found evidence that, although all 5 foundations are present across cultures, their weight can differ very sharply between cultural groups in accordance with the predictions of functionalist models of morality. The moral intuitions of western liberals, for example, derive in an unusually narrow manner from only two of the five Foundations (Care and Fairness) as opposed to a more balanced representation of all 5 found among conservative westerners or in non-western cultures like rural India (Haidt, 2007).

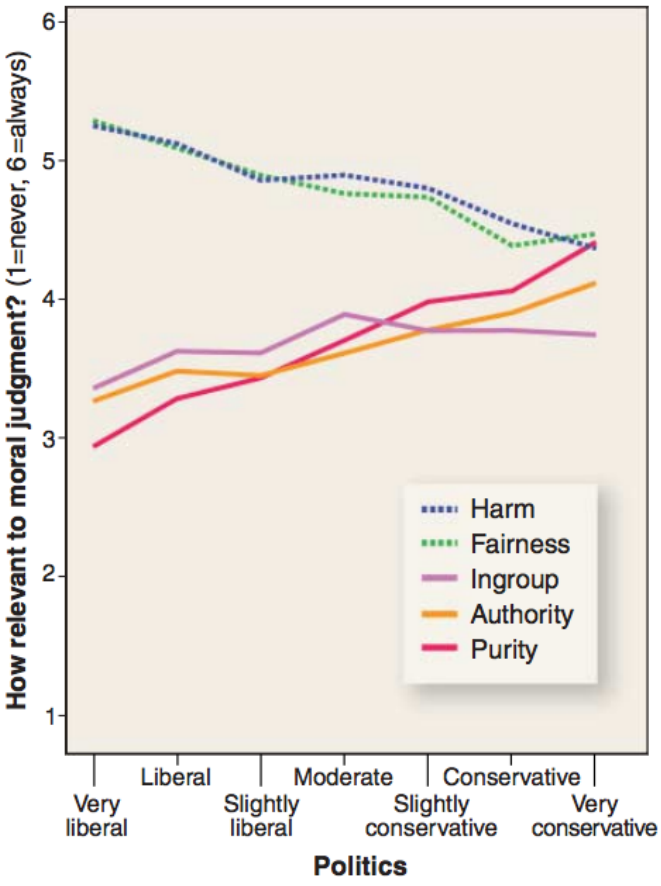


Figure 1. Judged relevance of each of Haidt's 5 Foundations across the political spectrum. Reproduced from Haidt, 2007.

MFT has a number of features which make it especially attractive among competing models. In addition to providing evolutionary accounts for the functions of its proposed foundations, MFT also performed better in comparative model-fitting by confirmatory factor analyses against alternatives with 1, 2, 3 and 4 domains, including Shweder's 3 ethics. (Graham et al., 2011). The theory's cultural-political neutrality is also a likely contributor to its appeal. Several of Kohlberg's intermediate stages of moral development are strongly reminiscent of foundations within MFT: the orientation to interpersonal conformity can be identified with the Loyalty foundation; orientation to authority and social-order maintenance are similar to the Respect foundation, and the utilitarian social contract orientation is arguably coincident with the Care and Fairness foundations. But Kohlberg's theory explicitly devalues Loyalty and Respect for their own sake, independent of utilitarian framing, as "conventional" stages, mere waypoints on the road to the more sophisticated moral orientations of Care and Fairness, which are closer to the universal, deontic ethics that occupy the top of the developmental hierarchy. In contrast, the foundations in MFT are all on adaptively level ground - each solves a distinct problem in the ancestral evolutionary environment, and social groups prioritize them by their own, culturally and geographically particular methods (Graham et al., 2011). Accordingly, MFT is more compatible than its antecedents with the wave of moral pluralism characterizing the academic arena of the early 21st century. Proponents of MFT also point to the differential weights on the foundations as a plausible account of the modern "culture wars" between liberal and conservative populations in many countries,

especially the United States (Graham, 2009).

Ballooning complexity and empirical shortcomings in moral models

Here it has to be noted that MFT has expanded twice since its original formulation. The first version had only four foundations, but within a short time from inception the authors differentiated the "Hierarchy" foundation into the current Loyalty and Respect. Recently, Haidt proposed a sixth foundation of Liberty/Oppression in light of new studies (2012). Taken together with previous work, these constitute continuous expansions from the 2-factor model of Kohlberg and Turiel, through Shweder's 3 ethics, to Haidt's 4, 5, and 6, and potentially beyond. With each refinement or elaboration of the putative structure of the moral domain, the number of subdivisions expands.

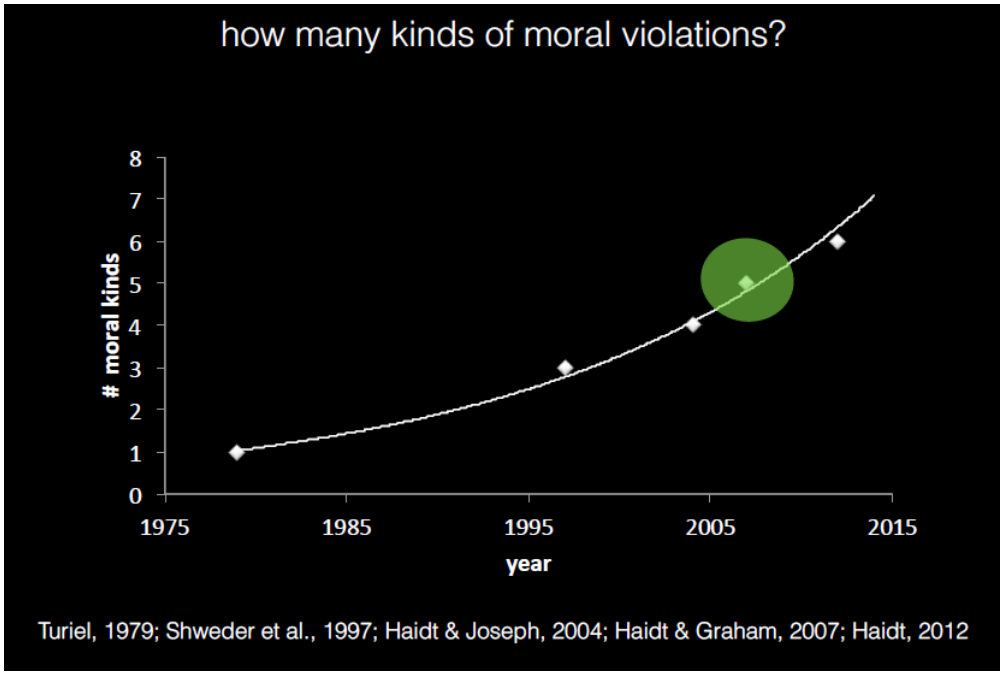


Figure 2. The number of factors in academically popular models of morality is increasing over time. This work focuses on a 5-factor model (highlighted).

But, having expanded the boundaries beyond WEIRD morality, is this new school now taking things too far by continuing to slice up the moral domain into ever-finer parts? No broad consensus has yet emerged as to which of these taxonomies are correct. In fact, Haidt's and Graham's own survey-based work showed ambiguous support for coarser versus finer divisions. Confirmatory factor analyses favored 5 factors (Graham et al., 2011), while exploratory methods found strong evidence for a coarser division into only two, superordinate factors, one made up of two "Individualizing" foundations (Harm and Fairness), so called because they are focused on effects on individuals, and one made up of three "Binding" foundations (Loyalty, Respect, and Sanctity), so called because they function at group level to bind people together into a greater whole (Graham et al., 2011).[2] The two superordinate groups are separable between social liberals and conservatives, individuals with psychopathy and healthy individuals, (Glenn, Iyer, Graham, Koleva & Haidt, 2009; Aharoni, Antonenko & Kiehl, 2011) and those operating under cognitive load versus typical function (Wright & Greene, 2013). It is important, then, to determine whether the added complexity of MFT is justified by a substantial increase in explanatory power, or whether the observed variance can be more satisfactorily accounted for by the simpler Individualizing/Binding taxonomy. While

[2] A word on the respective uses of exploratory versus confirmatory analyses: Confirmatory analyses are appropriate in later stages of model development for verification that previously established model scales correspond with observed responses. They should not be treated as direct support that a given map is parsimonious and exhaustive with respect to assaying a particular theory; for this purpose, exploratory, model-agnostic analyses are more appropriate.

larger numbers of factors this may lead to more intuitively compelling moral categories, it also detracts from parsimony. In addition, the methods by which investigators arrived at their models have largely been intuitive: lists of moral violations were generated by the researchers themselves, then grouped by those same researchers according to subjective criteria. Such methods are vulnerable to researcher biases. This may in part account for the trend of increasing complexity and lack of consensus.

In addition, work done so far on characterizing the structure of the moral domain has been mostly based on behavioral data. Only two major studies to date have investigated moral kinds as they are implemented in brain function and structure. One of these was limited to a single model of the moral domain comprising the 3 putative factors of harm, dishonesty, and disgust (Parkinson et al., 2011), corresponding to two of Shweder's three Ethics (autonomy and divinity) and three of Haidt's initial five Foundations (harm, fairness and purity). Based on fMRI findings of distinct neural systems correlated with the 3 transgression types, Parkinson et al. argued against the concept of a unified, monolithic human moral faculty (2011). The other work spanned 5 basic and 2 superordinate factors of Haidt's Moral Foundations theory (2007); however, the latter study investigated white matter volumes and, therefore, is only of indirect value in making functional associations (Lewis, Kanai, Bates & Rees, 2012). Beyond simply counting against unitary models of morality, these works can be taken as partial support for more complex models. Nonetheless, until recently, no work had been done explicitly to address either Shweder's or Haidt's frameworks in terms of number of functional brain networks, nor on whether and how moral *concepts* sort into groups which are similarly processed by the brain.

Exploratory, functional imaging analyses to probe neural correlates of moral domain structure

Alek Chakroff and I have worked to address these gaps in the research by conducting agnostic analyses searching for natural structure in fMRI data captured from subjects as they read from a list of moral transgressions. We have conducted exploratory data reduction analyses, including dimension reduction (e.g., principal components analysis [PCA]), on moral judgments and explicit similarity judgments of stimuli, and we have also employed searchlight-based Representational Similarity Analysis (RSA) to compare the activation patterns for each of our moral stimuli on the basis of derived representational matrices (Kemp & Tenenbaum, 2008; Kriegskorte, Goebel & Bandettini, 2006; Kriegskorte, Mur & Bandettini, 2008, Conolly et al., 2012). In addition to clustering voxels within the brain which co-activate in similar patterns in response to stimuli (functional units), they also allow for clustering of *stimuli* which elicit similar patterns of activation (Lashkari, Sridharan & Golland, 2010; Lashkari, Vul, Kanwisher & Golland, 2010) — in other words, stimulus categories, or, for moral stimuli, domains.

This has allowed us to test whether patterns of brain activity naturally sort along expected domain lines, and, if so, how many such domains exist and whether their boundaries correspond to previous theories based in cultural and evolutionary psychology. We initially anticipated that moral stimuli would group into either 3 domains, as suggested by Shweder, or 5, as suggested by Haidt. Among psychologists, MFT is presently the most popular and widely-held model, in part because it holds

several advantages over competitors: MFT provides plausible evolutionary accounts for the origins of its moral categories and confirmatory factor analyses have shown MFT to be a better fit to behavioral data than other models (Graham et al., 2011). Furthermore, MFT is the most complex and inclusive of the hypothetical moral structures under consideration; it can, in fact, be considered an elaboration, rather than a simple competitor, to Shweder's 3 ethics. However, early clustering results have defied expectations, instead appearing to support a 2-factor structure suggestive of the Individualizing/Binding super-foundation division, where one factor favors violations generally agreed to be wrong, such as physical harms and uncooperative behaviors, and the other favors violations generally judged wrong by social conservatives, such as sexual deviance, intoxication, or violations of social conformity. Confirmation of a 2-factor moral domain structure would be consistent with previous work (Haidt, 2007) while enjoying the substantial benefit of a simpler theoretical framework. However, because of the tendency of exploratory analyses to suffer from issues of low statistical power in comparison to confirmatory analyses of comparable scale, it is difficult to assert on these grounds alone (without expanding the design to thousands of subjects) that we are discovering the true structure rather than only the coarsest divisions.

The present work follows up on the question of the structure of the moral domain, including the behaviors it contains and the way in which those behaviors are neurocognitively interrelated, through the use of more traditional, model-based fMRI analyses to help resolve previous ambiguities. These analyses feature several improvements over previous methods and take Haidt and Joseph's 5-factor MFT as their

implicit hypothesis (Haidt & Joseph, 2004)[3]. Because the factors of MFT were elaborated largely by subdividing those proposed in previous models such as Shweder's 3 ethics, the simpler models can be indirectly addressed by factor aggregation. Another improvement over previous methods is the fact that the regressors for General Linear Model (GLM) analyses on fMRI scans are based on time-courses of moral stimuli which are both generated and sorted into the 5 categories of MFT by survey participants, rather than by the researcher, thereby avoiding the potential for researcher bias in stimulus generation. In this method, the moral domain is considered to consist of whatever people reliably tend to *call* moral, and functional correlates are then drawn to this set.

Previous work furnishes a set of brain regions expected to play a part in general morality-processing based on putative associations with morality-linked functions:

Function	**Associated regions**
Self-reference	• Medial prefrontal cortex (mPFC) (Northoff & Bermpohl, 2004; Whitfield-Gabrieli et al., 2011)

[3] As of this writing, Haidt has proposed a 6-factor elaboration of MFT introducing a new "Liberty/Oppression foundation ("Moral Foundations," n.d., para. 2-4). However, to date the majority of relevant work considering MFT has considered the 5-factor version. Furthermore, a discovered lack of support for the 5-factor version would make finding support for the more complex 6-factor version unlikely, while if support were found for 5 factors, the 6-factor version could be tested in follow-up work. For these reasons, I have limited myself to the 5-factor model.

	• Dorsomedial prefrontal cortex (dmPFC) (Whitfield-Gabrieli et al., 2011) • Posterior cingulate cortex (PCC) (Whitfield-Gabrieli et al., 2011) • Precuneus (Cavanna & Trimble, 2006)
Emotional valuation of stimuli	• vmPFC (Croft, 2009; Paulus & Frank, 2003) • Amygdala (Cunningham, Raye, & Johnson, 2004)
Attribution of enduring traits (such as trustworthiness and criminal responsibility) to others	• mPFC (Van Overwalle, 2009; Buckholtz et al., 2008) • Amygdala (Stanley et al., 2012) • Striatum (Buckholtz et al., 2008) • PCC (Buckholtz et al., 2008)
Social semantic categorization (stereotyping)	• mPFC (Contreras, Banaji, & Mitchell, 2012) • ACC (Contreras, Banaji, & Mitchell, 2012) • PCC (Contreras, Banaji, & Mitchell, 2012)

	• TPJ (Contreras, Banaji, & Mitchell, 2012)
Cognitive control and conflict resolution, especially in decision-making situations characterized by high risk, ambiguity or ambivalence	• Anterior cingulate cortex (ACC) (Botvinick, Braver, Barch, Carter & Cohen, 2001; MacDonald, Cohen, Stenger & Carter, 2000; Cunningham, Raye & Johnson, 2004) • Dorsolateral prefrontal cortex (dlPFC) (MacDonald, Cohen, Stenger & Carter, 2000; Nathaniel-James & Frith, 2001; Buckholtz et al., 2008) • Ventromedial prefrontal cortex (vmPFC) (Greene & Cohen, 2004; Fellows & Farah, 2007) • Frontal pole (FP) (Cunningham, Raye & Johnson, 2004) • Orbitofrontal cortex (OFC) (Cunningham, Raye & Johnson, 2004)
Control of attention and learning	• PCC (Pearson, Heilbronner, Barack, Hayden & Platt, 2011)

	• Inferior parietal lobule (IPL) (Singh-Curry & Husain, 2009)
Motor comprehension for understanding the intent of physical actions	• Supplementary motor area (SMA) (Nachev, Kennard & Husain, 2008)

Other regions may be tentatively associated with functions which are less general and plausibly linked to one of the proposed domain divisions of MFT:

Proposed moral factor	**Corresponding functions**	**Associated regions**
"Individualizing" (super-foundation including Harm and Fairness)	Prosocial behaviors	• dmPFC (Waytz, Zaki & Mitchell, 2012) • PFC (FeldmanHall, 2012)
"Binding" foundations (respect, authority, and, especially, purity)	Mentalizing/Theory-of-mind	• Temporoparietal junction (TPJ) (Young & Saxe, 2008; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Van Overwalle, 2009;

		Saxe, 2010) • mPFC (Mitchell, Banaji & Macrae, 2005; Young & Saxe, 2008) • Superior temporal sulcus (STS) (Redcay, 2008) • Precuneus (Young & Saxe, 2008) • Temporal pole [TP] (Olson, Plotzker, & Ezzyat, 2007)
	Disgust-related emotions	• Frontal/temporal regions including the insula (Moll et al., 2005)

The Individualizing super-foundation itself has also been positively associated with the volume of the lateral dmPFC and negatively with that of the precuneus, while the Binding super-foundation has been positively associated with the volume of the bilateral subcallosal gyrus and, non-significantly, with the volume of the left anterior insula (Lewis, Kanai, Bates, & Rees, 2012).

Other sets of regions have been linked to the polarity and intensity of political beliefs (ACC, mPFC, and TPJ with liberalism (Kanai, Feilden, Firth, & Rees, 2011; Zamboni et al. 2012), amygdala and dlPFC with conservativism, ventral striatum and PCC with radicalism in either pole (Zamboni et al., 2012)), and these associations may be taken as transitive support for their role in moral processing insofar as political attitudes are motivated by considerations of morality.

Chapter II

Materials and Methods

Participants

281 participants for Survey Group A were recruited through Amazon Mechanical Turk (www.mturk.com). Demographic information was collected from 200 of these participants. 47% were female. 23 countries of origin were represented, with the United States constituting 15% and India constituting 65%. A 7-point Likert scale ranging from "Very Liberal" to "Very Conservative" was used to gauge political orientation, and survey-takers had a good spread of orientations (each scale point was selected by at least 10 participants).

500 participants for Survey Group B recruited through Amazon Mechanical Turk. Demographic information was collected from all participants. A 7-point Likert scale ranging from "Very Liberal" to "Very Conservative" was used to gauge political orientation. 52 out of 500 participants did not complete the survey.

1096 participants for Survey Group C were recruited through Amazon Mechanical Turk.

100 participants for each of Survey Groups D and E were recruited through Amazon Mechanical Turk. Demographic information was collected from all participants. A 7-point Likert scale ranging from "Very Liberal" to "Very Conservative" was used to gauge political orientation.

All survey-takers were paid the standard Mechanical Turk sum for short tasks, $0.10.

For collection of fMRI scans, 18 adult, right-handed native English-speakers with normal or corrected-normal vision and no history of psychiatric or neurological issues were recruited to Scan Group A.

In a replication study, 20 adult, right-handed native English-speakers with normal or corrected-normal vision and no history of psychiatric or neurological issues were recruited to Scan Group B.

All subjects in scan groups gave written informed consent in accordance with the guidelines set by the Harvard University institutional review board. Participants were paid $60.00.

Measures

Moral stimulus collection

To generate moral stimuli, participants in Survey Group A were asked to "List as many moral transgressions as you can," and to describe each transgression with a brief passage. A list of 1,491 responses was collected, with an average per-participant response rate of 5.3.

The resulting list was edited to remove items which were redundant with one another, insufficiently elaborated (single-word responses like "theft"), vague or nonsensical (phrases like "evil" or "MY WORK ASK ANY MONEY") by three raters, including Alek Chakroff and two research assistants. Items were excluded based on vagueness or nonsensicalness if two out of three raters agreed to do so. Redundancies were resolved by combining less specific items into more specific items; for instance,

"murder" would be combined into the more specific "murdering your spouse." All remaining items were reworded into gerund form for consistency.

The resulting list was combined with a set of moral transgressions from the Moral Foundations Questionnaire (Graham et al., 2011) and 40 morally neutral items (e.g. "Going for a walk in the park") for a total set of 400 unique moral stimuli, 360 of which referred to a diverse range of transgression types. (See appendix for a subset of these transgressions.)

Stimulus rating and categorization

Participants in Survey Group B were each assigned a random subset of 90 moral transgressions from the full set of 360. Participants were asked to rate each violation on a 7-point Likert scale ranging from "Not at all wrong" to "Extremely Wrong." 98,640 transgression pairs were rated, with a per-pair average of 24.4 ratings.

A subset of 100 items out of the full 400 (90 of the 360 transgressions and 10 of the 40 neutral items) were categorized into the 5 foundations of MFT (Graham et al., 2011) by participants in Survey Group D. A different subset of 100 items (90 new transgressions and the same 10 neutral items) were categorized by Survey Group E. Participants in groups D and E were given definitions for the foundations based on Haidt's MFT definitions, summarized by myself for brevity and reframed in the exclusive negative (e.g. "Harm" instead of "Care/Harm") in accordance with the fact that non-neutral stimuli were transgressions of, rather than observances of, moral rules:

> ***Harm*** *refers to violations of an individual's material wellbeing. In these cases an action is wrong because it directly hurts another individual. To decide if an action is Harm-related, you*

think about things like injury, death, endangerment, impoverishment, and deprivation.

Unfairness *refers to violations of an individual's rights or freedoms. In these cases an action is wrong because it infringes upon his/her rights or freedoms as an individual. To decide if an action is Unfairness-related, you think about things like rights, justice, freedom, fairness, and the importance of individual choice and liberty.*

Disloyalty *refers to violations of the community. In these cases an action is wrong because a person fails to place the welfare of his or her social group above personal interest. To decide if an action is Disloyalty-related, you think about things like loyalty, group honor, and the preservation of the community.*

Disrespect *refers to violations of the social order. In these cases an action is wrong because a person fails to observe his or her place within a hierarchy. To decide if an action is Disrespect-related, you think about things like duty, role-obligation, interdependence and respect for authority.*

Impurity *refers to violations of purity and the divine. In these cases a person disrespects the sacredness of God, or causes impurity or degradation to himself/herself, or to others. To decide if an action is Impurity-related, you think about things like sin, the natural order of things, sanctity, and the protection of the soul or the world from degradation and spiritual defilement.*

Following the presentation of definitions, participants were asked to classify each of the items into one of 6 bins, one for each of the 5 foundations and one "not sure" bin. Within each bin, participants were instructed to arrange the items in rank order of fit to the foundation.

Procedures

Collection of fMRI data

Data was acquired with a Siemens 3T TimTrio scanner with a 32-channel headcoil at Harvard University Center for Brain Science. Prior to functional scans, a high-resolution, whole-brain structural scan (1 mm isotropic voxel MPRAGE) was acquired. Functional scans were acquired in 33 axial slices, interleaved, parallel to the AC-PC line, using an EPI pulse sequence with a TR of 2000 ms, a TE of 30 ms, a flip angle of 85, an FOV of 216 mm, and 3.0 mm isotropic voxels.

Scans were carried out as subjects in Scan Group A performed a judgment task on the 100 moral stimuli from Survey Group B on the previous classification task. The stimuli were presented using an Apple iMac running MATLAB and the Psychophysics Toolbox (Brainard, 1997) projecting onto a screen at the head of the magnet bore. Each of the 100 stimuli was presented 4 times, in randomized order, for a total of 400 presentations arranged into 16 runs of 100 prompts each. (Subject 4 was the only exception, with 14 runs instead.) Presentation time was 4 seconds, followed by a 10-second fixation period. Subjects were asked to judge the stimuli by degree of moral wrongness on a scale of 1 (“Not at all wrong”) to 5 (“Extremely wrong”), indicating with a button box held in the right hand.

In a replication study, subjects in Scan Group B performed the same task with the exception that a different subset of 90 non-neutral moral stimuli, categorized by Survey Group E, were presented. Data from Subject 1 of this replication group did not go on to

further analysis due to excessive head motion.

MATLAB and the SPM8 software were used for image preprocessing (Friston et al., 1995). The first four volumes were removed to allow for T_1 equilibration. Motion-correction was applied across and within runs. Images were spatially normalized to Montreal Neurological Institute (MNI) standard anatomical space and spatially smoothed with a 3 mm full-width half-maximum Gaussian kernel.

Chapter III

Results

GLM estimation

For the purposes of regressor generation, the modal classification of each item in the classification survey, excluding "not sure" classifications, was assigned as that item's primary foundation. Item assignments were not even across the foundations (Harm n=28, Unfairness n=32, Disloyalty n=12, Disrespect n=8, Impurity n=10). However, the subsequent results do not significantly change when standardizing the stimulus numbers across domains by randomly subsampling larger domains (e.g., Unfairness).

For Scan Group A, two sets of regressors were generated from the item classifications. The first set, referred to as "Binary," were constructed for each foundation by convolving a boxcar function for the onset of stimuli corresponding primarily to that foundation (1 when the foundation is present, 0 otherwise) with a standard hemodynamic response function. The second set, referred to as "Continuous," were constructed using a different boxcar function with values equal to the normalized fraction of the time that the stimulus was assigned to the given foundation versus all assignments (excluding "not sure" assignments). Accordingly, the Binary regressors reflect only the most characteristic foundation at each time point, while the Continuous regressors are allowed to divide up among multiple foundations at the same time if the stimulus was not cleanly categorized into only one. For instance, a stimulus categorized as harmful by 60% of survey participants, unfair by 35% and anti-authority by 5% would be modeled in three different regressors, multiplying the hemodynamic response function by 0.6, 0.35, and

0.5, respectively. For the replication with Scan Group B, only Continuous regressors were constructed. For both sets of regressors, neutral items were placed on a single regressor of no interest and, along with regressors of no interest for linear drift and between-run transitions, were not analyzed further.

General Linear Model (GLM) estimation was performed on the fMRI scan data with both the Binary and Continuous regressors using MATLAB and the SPM8 software package.

Between-foundation contrasts

For each subject, ten contrasts were computed in the MATLAB environment on SPM8 software, one for each possible pairwise comparison between the GLM images for two out of the five foundations. These contrasts were then combined across participants using group-level random effects analysis.

Conjunction analyses on between-foundation contrasts

In order to determine whether foundations within the 2 hypothesized super-foundations (Individualizing and Binding) behaved similarly, conjunction analyses were performed on the results of the group-level random effects analysis from the between-foundation contrasts. If the 2 super-foundations are functionally meaningful, then foundations should contrast in a similar way (in terms of overlap in significant voxels) against other foundations when both contrasts have the same relationship to super-foundation boundaries. For example, because the Care and Fairness foundations share the Individualizing super-foundation while the Respect and Purity foundations share the

Binding super-foundation, the contrast Care > Purity would be expected to be similar to the contrast Fairness > Purity (because each contrast contains one Individualizing and one Binding foundation) and dissimilar to the contrast Respect > Purity (because both are Binding foundations). However, Respect > Purity may be expected to be similar to another conjunction of intra-Binding foundations, such Loyalty > Respect.

Input images were thresholded using t-values corresponding to a p-value of 0.05 and then multiplied using the imcalc function in the SPM8 software package, resulting in output images identifying only voxels of overlapping contrast. Percentage overlap between each pair contrasts was then calculated as the number of voxels common to both contrasts divided by the larger of the two voxel counts, with the positive and negative regions of each conjunction handled separately and then summed to produce absolute-value conjunctions.

Results of between-foundation contrasts

Here I report results of the contrasts where both Binary- and Continuous-regressor methods agreed on activations within the regions of interest.

For contrasts of the Individualizing foundations against one another and against the Binding Foundations, stimuli identified with the Care foundation generally corresponded to lower activations in areas associated with cognitive control and the resolution of ambiguity and conflict, including dlPFC and both ACC and PCC, than did stimuli associated with all other foundations (including Fairness). This distinction was weaker in the contrast of Care vs Purity, where different regions of the dlPFC were favorably activated in either condition, and in which the finding of lower activation in

ACC failed to replicate in Scan Group B. Care was also associated with higher activation than Loyalty in STS, an area associated with mentalizing.

In addition to higher cognitive-control-associated activations in comparison to Care, stimuli identified with the Fairness foundation corresponded to higher activations in regions associated with cognitive control/conflict resolution (dlPFC) and understanding the cognitive and motor intentions of others (precuneus, SMA) than did the Binding foundations. It should be noted, however, that areas putatively involved in cognitive control besides the dlPFC, including the cingulate cortex (MacDonald, Cohen, Stenger & Carter, 2000), are less associated with Fairness-foundation stimuli than with stimuli from the Binding foundations, Loyalty, Respect, and Purity. The striatum, a region associated with the attribution of persistent valuations to others (Buckholtz et al., 2008), was less active in the Fairness condition than in the Purity condition. All of the cingulate- and striatum-related differntial activations in Fairness failed to replicate in Scan Group B.

For contrasts among the three Binding foundations, stimuli identified with the Loyalty foundation corresponded to higher activations than Respect in an area associated with understanding motor intent (SMA), while Respect corresponded with higher activations in a cognitive-control-associated region (dlPFC) than did Loyalty; this latter finding failed to replicate in Scan Group B. Both Loyalty and Respect, when contrasted to Purity, had higher activations in regions associated with self-reference, attribution of enduring traits and stereotyping (PCC, mPFC), while Respect also exhibited higher activations in regions associated with cognitive control (dlPFC) and mentalizing (precuneus); out of these differential activations, the only one to replicate in Scan Group B was the finding of higher mPFC activation in the Respect versus Purity condition.

Purity stimuli corresponded to higher activation than Respect in an area associated with attribution of enduring traits and stereotyping (PCC), which did replicate in Scan Group B.

Second foundation in contrast	First foundation in contrast: Care	Fairness	Loyalty	Respect	Purity
Care		**Right inferior frontal gyrus** Subgyral temporal lobe/ hippocampus **Occipital gyrus**	Middle temporal gyrus Supplementary motor area Cingulate Cuneus Lingual/occipital gyrus	**Middle frontal gyrus** Inferior parietal lobule Precuneus Hippocampus	**Inferior frontal gyrus** Insula **Cingulate**
Fairness	Medial frontal gyrus Left inferior frontal gyrus Temporal/insula Inferior parietal lobule		Middle frontal gyrus Insula Superior temporal gyrus **Cingulate** Fusiform gyrus	Middle frontal gyrus Medial frontal gyrus **Postcentral gyrus** **Cingulate**	Middle frontal gyrus **Inferior parietal lobule** **Cingulate**/paracentral lobule **Lentiform nucleus**/ Globus pallidus
Loyalty	**Superior temporal gyrus** Postcentral gyrus Inferior parietal lobule	**Inferior frontal gyrus** Precentral gyrus Middle temporal gyrus Angular gyrus		**Superior frontal gyrus** **Middle frontal gyrus** **Postcentral gyrus** Inferior parietal lobule Superior temporal gyrus Cingulate Precuneus	Medial frontal gyrus Inferior frontal gyrus Supplementary motor area Inferior parietal lobule Cingulate Putamen/lentiform nucleus
Respect	Posterior cingulate **Lingual gyrus** **Thalamus**	**Inferior frontal gyrus** **Posterior cingulate** **Precuneus** **Lingual gyrus**	**Precentral gyrus** Posterior cingulate Cuneus Fusiform gyrus **Lingual gyrus** **Thalamus** Cerebellum		Inferior frontal gyrus Insula Middle temporal gyrus **Posterior cingulate** Fusiform gyrus
Purity	Medial frontal gyrus **Inferior frontal gyrus** Precentral gyrus Cuneus **Lingual gyrus**	Middle frontal gyrus **Precentral gyrus** **Lingual gyrus** **Cerebellum**	Cingulate **Posterior cingulate** Lingual/occipital gyrus	Superior frontal gyrus **Middle frontal gyrus** **Medial frontal gyrus** **Precuneus**	

Table 1a. Brain regions with greatest differences in mean BOLD activation (count of voxels passing significance) when using Binary regressors derived from Scan Group A. Regions listed are those in which the first foundation in the contrast is associated with more activity than the second foundation. Regions in bold were also detected in the contrasts from Continuous regressors.

Second foundation in contrast	First foundation in contrast: Care	Fairness	Loyalty	Respect	Purity
Care		Middle frontal gyrus **Inferior frontal gyrus** Precuneus Lingual/**occipital gyrus**	Frontal lobe	**Middle frontal gyrus** Medial frontal gyrus Angular gyrus	Inferior parietal lobule **Inferior frontal gyrus** **Cingulate** Cerebellum
Fairness	Superior temporal gyrus Postcentral gyrus Superior frontal gyrus		Superior frontal gyrus Postcentral gyrus **Cingulate** Putamen/lentiform nucleus	Superior frontal gyrus **Postcentral gyrus** Superior temporal gyrus **Cingulate**	Supplementary motor area **Inferior parietal lobule** **Cingulate** Putamen/**lentiform nucleus**
Loyalty	Superior frontal gyrus **Superior temporal gyrus** Posterior cingulate	**Inferior frontal gyrus** Superior temporal gyrus Lingual gyrus Thalamus		**Superior frontal gyrus** **Middle frontal gyrus** **Postcentral gyrus** Insula Posterior cingulate	Postcentral gyrus Supramarginal gyrus Superior temporal gyrus
Respect	Superior temporal gyrus Anterior cingulate Precuneus **Lingual gyrus** Fusiform gyrus **Thalamus**	**Inferior frontal gyrus** **Lingual gyrus** Thalamus	Middle frontal gyrus **Precentral gyrus** Precuneus/cuneus **Lingual gyrus** **Thalamus**		Medial frontal gyrus Postcentral gyrus Cingulate **Posterior cingulate** Cerebellum
Purity	**Inferior frontal gyrus** Superior temporal gyrus **Lingual gyrus** Thalamus	Superior frontal gyrus Medial frontal gyrus **Precentral gyrus** Middle temporal gyrus Inferior temporal gyrus **Lingual gyrus** **Cerebellum**	Superior frontal gyrus Middle frontal gyrus Medial frontal gyrus Precentral gyrus Middle temporal gyrus **Posterior cingulate** Cuneus	**Middle frontal gyrus** **Medial frontal gyrus** **Precuneus**	

Table 1b. Brain regions with greatest differences in mean BOLD activation (count of voxels passing significance) when using Continuous regressors from Scan Group A.

Regions listed are those in which the first foundation in the contrast is associated with more activity than the second foundation. Regions in bold were also detected in the contrasts from Binary regressors.

Second foundation in contrast	First foundation in contrast: Care	Fairness	Loyalty	Respect	Purity
Care		Middle frontal gyrus **Inferior frontal gyrus** Precuneus Lingual/**occipital gyrus**		**Middle frontal gyrus**	Inferior parietal lobule **Inferior frontal gyrus** Cerebellum
Fairness			Postcentral gyrus Putamen/lentiform nucleus	Superior frontal gyrus **Postcentral gyrus** Superior temporal gyrus	**Inferior parietal lobule**
Loyalty	Superior frontal gyrus **Superior temporal gyrus**	Lingual gyrus		**Superior frontal gyrus** **Postcentral gyrus**	Postcentral gyrus Superior temporal gyrus
Respect	Superior temporal gyrus	**Inferior frontal gyrus** **Lingual gyrus**	Middle frontal gyrus **Precentral gyrus** Precuneus		Medial frontal gyrus **Posterior cingulate**
Purity	**Inferior frontal gyrus** Superior temporal gyrus **Lingual gyrus** Thalamus	**Precentral gyrus** **Lingual gyrus** **Cerebellum**	Superior frontal gyrus Middle frontal gyrus Medial frontal gyrus Precentral gyrus Middle temporal gyrus Cuneus	**Medial frontal gyrus**	

Table 1c. Brain regions with greatest differences in mean BOLD activation (count of voxels passing significance) when using Continuous regressors derived from the replication in Scan Group B. Regions listed are those in which the first foundation in the contrast is associated with more activity than the second foundation. Regions in bold were also detected in both Binary and Continuous contrasts from Scan Group A.

The mean difference in absolute differential BOLD activation in the between-subject analyses of the contrasts from Continuous regressors derived from Scan Group A, defined as the number of individual voxels passing a significance level of $p = 0.5$ in each condition, can be interpreted as an estimate of the absolute difference in activations between the two conditions of each contrast. Mean difference in absolute BOLD activation was greatest between the Care and Respect conditions (25536 voxels passing significance), followed by Care/Purity (23688), Care/Fairness (20605), Fairness/Purity (18936), Care/Loyalty (17667), Fairness/Loyalty (13998), Loyalty/Respect (13363), Loyalty/Purity (12598), Fairness/Respect (11944), and Respect/Purity (9337).

Mean difference in only positive differential BOLD activation was greatest between the Care and Fairness conditions (9526), followed by Loyalty/Purity (6873), Fairness/Purity (6313), Care/Purity (5806), Respect/Purity (5647), Loyalty/Respect (4626), Fairness/Loyalty (3980), Care/Loyalty (3180), Fairness/Respect (3031), and Care/Authority (2603).

Mean difference in only negative differential BOLD activation was greatest between the Care and Respect conditions (22933), followed by Care/Purity (17882), Care/Loyalty (14487), Fairness/Purity (12623), Care/Fairness (11079), Fairness/Loyalty (10018), Fairness/Respect (8913), Loyalty/Respect (8737), Loyalty/Purity (5725), and Respect/Purity (3690).

The replication on data from Scan Group B yielded considerably different results. Mean difference in absolute differential BOLD activation was greatest between the Purity and Respect conditions (43055), followed by Care/Purity (40842), Loyalty/Purity (33675), Fairness/Purity (22116), Care/Fairness (21086), Care/Respect (20629),

Fairness/Respect (19024), Care/Loyalty (16886), Loyalty/Respect (13857), and Fairness/Loyalty (11457).

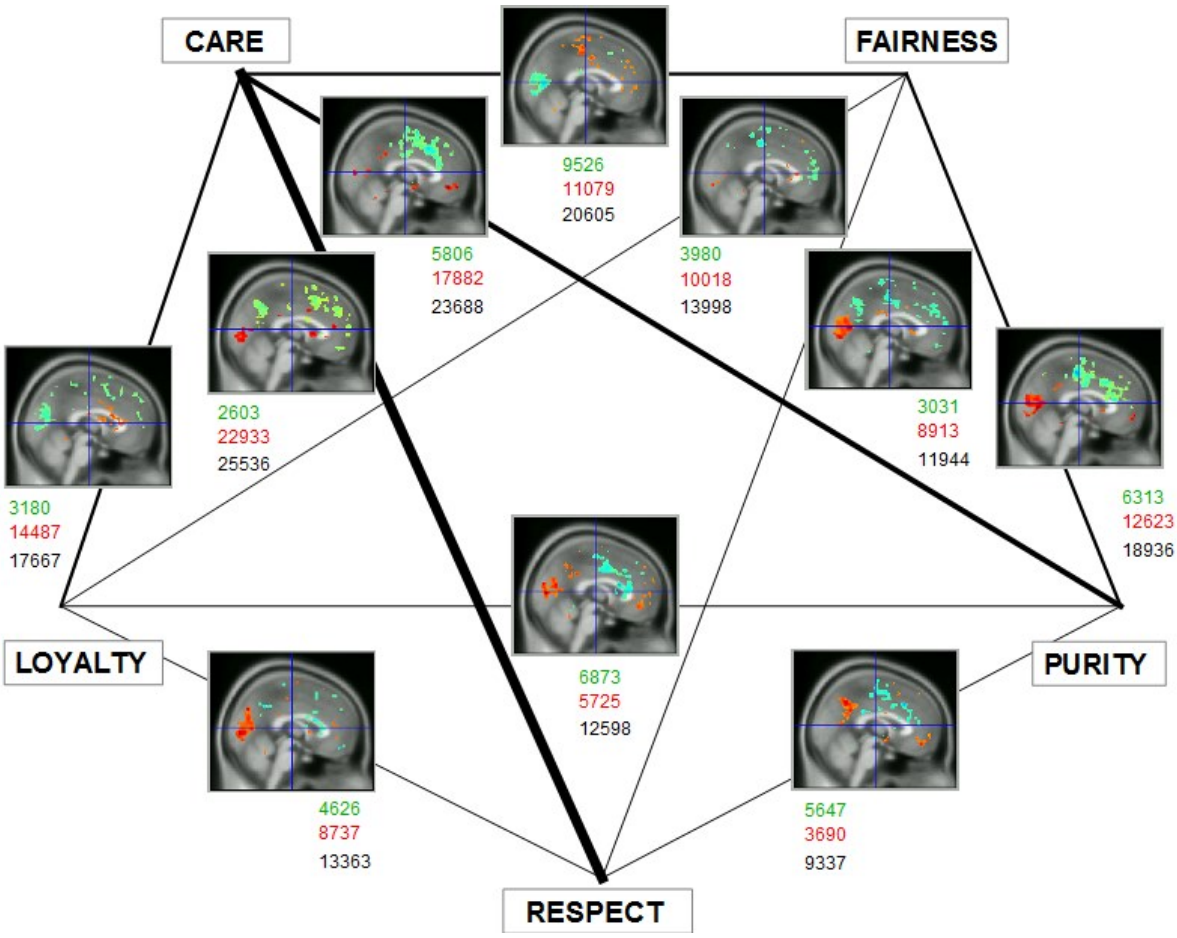
Mean difference in only positive differential BOLD activation for the replication was also considerably different, being greatest between the Respect and Purity conditions (38898), followed by Care/Purity (32576), Loyalty/Purity (30014), Fairness/Purity (17735), Care/Fairness (8478), Fairmess/Respect (7177), Loyalty/Respect (5985), Care/Loyalty (4846), Fairness/Loyalty (3987), and Care/Respect (3105).

Mean difference in only negative differential BOLD activation for the replication more closely resembled results for Scan Group A, being greatest between the Care and Respect conditions (17524), followed by Care/Fairness (12608), Care/Loyalty (12040), Fairness/Respect (11847), Care/Purity (8266), Respect/Loyalty (7872), Fairness/Loyalty (7470), Fairness/Purity (4381), Respect/Purity (4157), and Loyalty/Purity (3661).

Figure 3. Group-level contrasts between each of the five moral foundation GLM images. Using a threshold of $p < .001$, uncorrected, with a cluster-size threshold of 5 voxels. Only contrasts from Continuous regressors are shown here.

3a. Scan Group A. Numbers indicate the number of voxels passing threshold for each contrast, across the whole brain (summed across opposing contrasts, such as C>F + F>C), with green indicating positive contrast, red indicating negative contrasts, and grey/black indicating absolute contrasts (positive + negative).

3b. Scan Group B. Numbers indicate the number of voxels passing threshold for each contrast, across the whole brain (summed across opposing contrasts, such as C>F + F>C), with green indicating positive contrast, red indicating negative contrasts, and grey/black indicating absolute contrasts (positive + negative).



3c. Images reflect the contrasts C>F, C>L, C>R, C>P, F>L, F>R, F>P, L>R, L>P, P>R for Scan Group A.

Numbers are the same as in 5a.

Results of conjunction analyses on between-foundation contrasts

Conjunctions of between-foundation contrasts from Scan Group A had generally low percentage overlaps in the 0-20% range, with no overlaps greater than 50%. As expected, conjunctions of contrasts within the Binding super-foundation showed relatively strong conjunctions (greater than or equal to 10% overlap by both Binary and Continuous regressors) when excluding conjunctions that placed the same foundation on opposite "sides" in each contrast (e.g. the conjunction of Loyalty versus Respect with Respect versus Purity). Conjunctions of contrasts between the foundations in the Individualizing and Binding super-foundations were weaker as a group. Care and Fairness were both found to behave relatively similarly to one another when contrasted against the foundations of the Binding super-foundation.

Interestingly, a large number of conjunctions where one contrast was between super-foundations and one was within superdomains showed relatively high overlap. Out of 18 such conjunctions, 9 were dissimilar (less than 10% overlap by both Binary and Continuous regressors) while the other 9 were similar (10% overlap or greater, by both Binary and Continuous regressors in 8 out of the 9 cases). In particular, Care and Fairness behaved similarly to Loyal in contrasts against Respect and Purity. Care and Fairness also behaved similarly to Respect against Purity.

The two Individualizing foundations (Care and Fairness) had a relatively high and unambiguous degree of similarity in their contrasts against all three foundations of the Binding super-foundation (Loyalty, Respect, and Purity). When ranking conjunctions of

contrasts from highest to lowest degree of overlap, such conjunctions of between-superfoundation contrasts were in the top 5 by both Binary and Continuous regressors. Fairness also showed a relatively high degree of internal consistency in its contrasts against the three Binding foundations.

Conjunctions of contrasts which did not straddle the super-foundation boundary (one or more of the contrasts was between foundations falling fully within either Individualizing or Binding) were generally weaker. However, Loyalty had relatively high conjunction overlap with both Care and Fairness when all three were contrasted against Purity. Among the relatively strong conjunctions of between-superfoundation contrasts mentioned above, those involving Loyalty were weakest. Taken together, this evidence may suggest that Loyalty fits less well into the Binding super-foundation, and more well into Individualizing, than expected.

On replicate analysis from Scan Group B the pattern of conjunctions differed, with two conjunctions having greater than 50% overlap. These involved contrasts of Care and either Loyalty or Respect against Purity, and were followed immediately by the conjunctions of Loyalty and Respect versus Purity and Care and Fairness versus Purity.

Figure 4. Overlap in significant voxels between pairs of contrasts. Only absolute overlaps are shown (conjunctions for positive and negative contrasts are summed). Numbers along the diagonal refer to the number of significant voxels in a single contrast. All other numbers refer to the overlap between two different contrasts relative to the contrast in the pair with the larger number of significant voxels, with 1 representing perfect overlap. Conjunctions of contrasts going in different "directions" (i.e. with Respect in the first position in contrast 1 and also the second position in contrast 2) all had conjunctions below 0.01 and are not shown. The interior, red-outlined rectangle bounds the set of conjunctions of only between-super-foundation contrasts (1 foundation of each of contrast is from the Individualizing foundations and the other is from the Binding foundations).

	Care / Fairness	Care / Loyalty	Care / Respect	Care / Purity	Fairness / Loyalty	Fairness / Respect	Fairness / Purity	Loyalty / Respect	Loyalty / Purity	Respect / Purity
Care / Fairness	16840									
Care / Loyalty	0.256	14476								
Care / Respect	0.096	0.094	20306							
Care / Purity	0.057	0.044	0.065	7031						
Fairness / Loyalty		0.249	0.044	0.023	8900					
Fairness / Respect		0.028	0.471	0.041	0.055	21890				
Fairness / Purity		0.011	0.056	0.278	0.056	0.147	11643			
Loyalty / Respect	0.022		0.387	0.043		0.491	0.109	21868		
Loyalty / Purity	0.011		0.053	0.195		0.105	0.371	0.2	12221	
Respect / Purity	0.032	0.027		0.11	0.03		0.124		0.104	9847

4a. Results from contrasts using Binary regressors on Scan Group A.

	Care / Fairness	Care / Loyalty	Care / Respect	Care / Purity	Fairness / Loyalty	Fairness / Respect	Fairness / Purity	Loyalty / Respect	Loyalty / Purity	Respect / Purity
Care / Fairness	20605									
Care / Loyalty	0.112	17667								
Care / Respect	0.05	0.209	25536							
Care / Purity	0.05	0.102	0.122	11994						
Fairness / Loyalty		0.302	0.123	0.069	13998					
Fairness / Respect		0.117	0.392	0.083	0.301	23688				
Fairness / Purity		0.096	0.151	0.319	0.25	0.353	18936			
Loyalty / Respect	0.022		0.204	0.054		0.182	0.102	13363		
Loyalty / Purity	0.045		0.05	0.342		0.081	0.242	0.205	12598	
Respect / Purity	0.039	0.019		0.3	0.019		0.136		0.254	9337

6b. Results from contrasts using Continuous regressors on Scan Group A.

	Care / Fairness	Care / Loyalty	Care / Respect	Care / Purity	Fairness / Loyalty	Fairness / Respect	Fairness / Purity	Loyalty / Respect	Loyalty / Purity	Respect / Purity
Care / Fairness	21086									
Care / Loyalty	0.2	16886								
Care / Respect	0.194	0.178	20629							
Care / Purity	0.11	0.088	0.081	40842						
Fairness / Loyalty		0.167	0.044	0.016	11457					
Fairness / Respect		0.051	0.226	0.026	0.249	19024				
Fairness / Purity		0.04	0.047	0.311	0.131	0.216	22116			
Loyalty / Respect	0.05		0.217	0.029		0.204	0.059	13857		
Loyalty / Purity	0.051		0.032	0.527		0.041	0.309	0.097	33675	
Respect / Purity	0.063	0.042		0.618	0.014		0.274		0.466	43055

4c. Results from contrasts using Continuous regressors derived from replication on Scan Group B.

Chapter IV

Discussion

Haidt's 5-factor hypothesis was used as a framework for regressor classification in this study. Distinct brain-activation maps for the 5 foundations are necessary to support Haidt's formulation. The contrast of activation maps for different moral foundation conditions examined here support previous findings (Parkinson et al., 2011) indicating that morality is neither a unified nor a centrally localized function of the brain, but instead a whole-brain phenomenon carried out by several distinct networks of widely-distributed brain regions. However, the purpose of this work is to narrow the range of plausible structures hidden within that variety. It was hypothesized that variations in patterns of BOLD activation between the five foundations should be greatest between, rather than within, the two sets defined by the Individualizing and Binding super-foundations. That is, stimuli associated with Care/Fairness on the one hand and Loyalty/Respect/Purity on the other, if those groupings are functionally meaningful, should show greater activation disparities when compared against one another than against the other members of their super-foundation.

Considering contrasts of the Individualizing foundations against one another and against the Binding Foundations, there is less common activation and regional overlap than expected within the Individualizing super-foundation. The larger distinction seems not to be between Care and Fairness as a group against the other three foundations, but between Care alone and everything else. As defined by the number of voxels in each contrast passing a significance level of $p > 0.05$, 4 out of 5 of the largest differences in

absolute significantly differential BOLD activations in the initial scan group occurred between the Care foundation and all other foundations, including Fairness. The Care foundation was associated with less activation in cognitive control and ambiguity-resolution-associated regions than were any of these other foundations (which may reflect a difference in Care-centered processing generally or, alternately, that the Care-related prompts themselves were simply less demanding in control terms). Considering the contrasts from continuous regressors alone, the STS, an area associated with mentalizing and understanding social intent, was preferentially activated in Care versus all other foundations. The generally lower activation in control-associated regions for the Care foundation must be qualified in the contrast of Care against Purity. In this particular case, different regions of the dlPFC are preferentially activated in each condition, and dlPFC was not identified as a region of significant contrast on replication.

In comparison, in the initial scan group, Fairness was very highly associated with dlPFC, a region identified with cognitive control and analytical/utilitarian calculations, though it was less associated with activation in cingulate cortex than were the Binding foundations (and, in fact, the Care foundation). The anterior cingulate has been posited by Botvinick et al. to be more concerned with conflict-detection, and the dlPFC with implementation of the corresponding control (2001); accordingly, these results may suggest that the Binding-foundation-associated stimuli more easily put the brain into a state of alert, while Fairness-associated stimuli engage deliberative processes more directly and with less alarm. There was also notable difference in absolute BOLD activations between Fairness and Purity. The striatum, which is associated with the attribution of enduring traits to stimuli (Buckholtz et al., 2008), shows higher activation

under the Purity condition than Fairness. These results in cingulate and striatum were not, however, repeated in the replicate analysis in the second scan group, so a high degree of caution must be applied to any conclusions suggested by them.

The Binding foundations showed relatively weak differences in absolute significantly-differential BOLD activations against one another, as expected. Loyalty was more associated with motor-intent-associated regions than Respect. Both were more associated with regions identified in cognitive control, attribution of enduring traits and stereotyping than was Purity, with the exception that Purity exceeded Respect in posterior cingulate activation. In both scan groups, Purity was found to have generally high contrast in positive activations against other groups, both within and between superfoundations, while Respect had notably high contrast in negative activation against Care.

In conjunctions of contrasts between the Individualizing and Binding super-foundations, Care and Fairness were both found to behave similarly to one another in contrasts against the foundations of the Binding super-foundation, relative to the generally low conjunction percent overlaps overall. These between-superfoundation conjunctions were weakest when Loyalty was one of the contrasted foundations. Conjunctions including those contrasts which did not straddle super-foundations were relatively weak, with one exception: Care and Fairness behaved similarly to Loyalty in contrasts against Purity.

Considering contrasts and conjunctions from the initial scan group together suggests that, while stimuli of both Care and Fairness produce distinct activation patterns from the Binding foundations, they are much less alike than hypothesized, with Fairness

associated much more with activations in areas related to cognitive control than is Care. There is also evidence that Care and Fairness may be more alike with Loyalty and Respect than previously thought and that Purity occupies a more outlying position even within the Binding group. Replication analyses do not necessarily undermine the observed distinction between the Individualizing foundations, though it is somewhat weaker, while the commonalities of contrasts involving any of the foundations versus Purity are much stronger on replication. These results may suggest a structure unlike those previously put forward, one in which both Care and Purity take distinct, outlying positions with Fairness, Loyalty and Respect in the intervening positions and more interrelated. This possibility stands somewhat at odds with previous conclusions by Alek Chakroff and myself (pointing to 2 factors) and Haidt (5 factors) while also failing to correspond cleanly to the 3 ethics of Shweder.

Research Limitations

A potential methodological drawback of this experimental design is in the collection of moral violation stimulus items. Survey participants were given the prompt "List as many moral transgressions as you can." The resulting list may be biased towards those moral violations which are easy to call to imagination, rather than broadly inclusive of all moral violations including those which are unusual but would be judged by many as morally wrong regardless of their peculiarity. As an example, none of the items concerned cannibalism, though cannibalism is widely viewed as a strong moral wrong. (Chakroff & Young, 2015; Russell & Giner-Sorolla, 2011) However, the use of "top-of-mind" violation items is argued for by the impracticality of presenting an exhaustive list

of possible violations to survey participants, and by the observation that, if moral intuitions serve practical social functions, (Haidt & Graham, 2007) then those moral violations which are easiest to imagine may also be those the participants consider most relevant to their own lives.

Another methodological limitation is in the construction of the continuous (foundation-weighted) GLM regressors. The foundation weight of each item is not independent of the weights for the other foundations since, once an item is sorted into a given foundation's bin, it is no longer available to be sorted into any of the other bins. This may lead to difficulties in interpretation. However, the analysis was carried out in this fashion because of the potential to yield results with high statistical power. Statistical power is essential here, as there is a risk that any discovered divisions in functional anatomy could reflect only the coarser part of the actual structure, with finer structure lying below the limits of detection. The replication study using a different subset of 90 non-neutral moral stimuli from the original set of 360 collected is included to help rule out such methodological failures due to low statistical power.

Another distinct possibility is that information contained in that finer structure of activations itself - that is to say, the fine-grained pattern of stimulus-associated BOLD signal, rather than the region-averaged intensity of BOLD signal. Further analyses should compare these patterns of brain activity for each of the five hypothesized moral foundations and the two hypothesized super-foundations; this question will be addressed through Representational Similarity Analysis (RSA) searchlight analyses (Kriegskorte, Goebel, & Bandettini, 2006; Kriegskorte, Mur, & Bandettini, 2008) in pending manuscript by Chakroff (2014) and are therefore not included in the scope of this work.

Implications

The findings of this work do not clearly support one of the previously advanced moral domain structures. While there is a strong observed polarity between the putative Care and Purity factors of MFT, the other 3 foundations occupy intermediate positions rather than the clean break between Individualizing and Binding foundations that was expected. This may indicate that our approach to morality and the brain requires some readjustment. It is plausible that previously observed moral factors aren't fundamental, but instead merely represent points in a continuous multi-factorial space defined by other qualities which are *not* necessarily moral, such as emotional salience, engagement of cognitive control or mentalizing. For example, moral cognition associated with the Care and Purity foundations differ sharply in their modulation by perceived intentionality: judgments of harm violations can be highly mitigated by considerations of intent, while purity violations are robust against such considerations (Young & Saxe, 2011). If this is the case, then future research would do well to focus on these underlying factors instead of the moral types themselves.

Mapping the divisions in human morality is a critical piece of the "new synthesis" in moral and social psychology as described by Haidt: in addition to refined stances on moral intuition versus reasoning, the social functionality of moral cognition, and its coevolution with cultural practice, he highlights the principle that "morality is about more than [the] harm and fairness" with which moral psychology has been preoccupied for the better part of the past century (Haidt, 2007). However complicated the actual structure of the moral domain may be, it is imperative that we solidify models of that structure for

both the specialist and lay communities during a time in which the neuropsychology of morality is assuming greater importance in the sociological, political, legal, and ethical fields. The proponents of MFT have suggested that moral orientations, implemented in the brain as distinct and differentially-weighted moral kinds, may be causal factors in the "culture wars" cutting across diverse communities worldwide (Graham, Haidt & Nosek 2009). A well-supported model might promote understanding between charged ideological poles, who currently lack a solid framework where two parties can come to different conclusions despite both approaching a moral question in good faith. If, in fact, morality ultimately reduces to a set of non-moral dimensions, then we would do well to focus conversations on these underlying considerations. Doing so could shift those conversations away from irreconcilable declarations of subjective moral "truths" and towards considerations of the differences in values between parties - and how, practically speaking, to best satisfy *everyone's* values, even when they do differ. In other words, mapping the moral domain in non-moral terms could open alternative paths around the historically obstructive rancor which so often paralyzes efforts at cooperation across our cultural divides.

Appendix: Subset of stimuli used in study

Two unmarried minors having sex.
Euthanasia and assisted suicide.
Consuming alcohol before age 18
Disobeying your parents.
Marrying a person of the same sex
Appearing nude in public.
Starting a business without planning ahead.
Driving a car without a valid inspection
Selling sexual favors for money.
Cursing at your parents.
Students misbehaving in class.
Presenting yourself insincerely.
Riding a jet ski in front of a fisherman, disrupting his spot.
Disobeying your teacher.
Being addicted to alcohol
Always criticizing your country or your workplace.
Watching obscene movies.
Being an Atheist.
Refusing to adopt an eco-friendly lifestyle.
Trespassing on someone's private property.
Purchasing cocaine for recreational use
Bumping into someone's car and not telling anybody.
Bribing a technician to get free cable.
Manipulating the outcome of a contest through violation of the rules.
Smoking in a non-smoking area.
Hiding a problem at work until the day you go on vacation.
Breaking the rules of your school.
Not paying income tax to the government
Becoming addicted to drugs.
Running a red light.
Allowing a cashier to give you too much change
Blindly obeying orders.
Using your company’s work car for private purposes
Being intolerant of different opinions than your own.
Forging another person's signature.
Talking about someone in their absence.
Misappropriating the funds entrusted to you
Adulterating food to save money.
Cheating at games when others play fair
Buying morally compromised items (blood diamonds, chocolate farmed by slaves)

Making fun of someone's culture.
Sleeping with a manager to get a promotion.
Spray-painting a public playground.
Buying friends dinner and submitting it as business expenses to work
Downloading digital music from the internet without paying
Allowing beggars to live in poverty.
Keeping friends only because they are useful.
Cheating customers in an online marketplace.
Having intercourse with an animal.
Bribing a police officer.
Judging others.
Cheating a friend out of their money.
Eating your pet dog after it has been hit and killed by a car
Polluting the air.
Exploiting the environment.
Forcing children to work for money.
Refusing to help a blind person
Exploiting the ignorant for political gain.
Forcing your political or religious views on someone else.
Filming somebody without their permission and publishing it.
Refusing to help someone get to the hospital.
Refusing to give water to a thirsty stranger.
Looting an ATM.
Exploiting the poor.
Becoming a terrorist against your own country.
Betraying the trust that someone has put in you
A Teacher spanking an unruly student.
Sexually harassing a coworker.
Looting a store.
Killing others in the name of religion.
Lying about a coworker to get them fired.
Teasing a girl about being fat.
Factory farming animals in inhumane conditions.
Killing someone for money.
Getting children addicted to smoking.
Abusing someone without any reason
Setting fire to someone's house.
Murdering your parents.
Refusing to hire someone because of their race.
Sending someone obscene e-mails against his or her wish.
Acting out of hatred of homosexuality.
Kidnapping a child to raise as your own.
Murdering your spouse.
Stealing someone's identity
Conducting medical trials on unwilling patients.

Working as a suicide bomber.
Intentionally infecting someone with AIDs.
Having sexual intercourse with a child.
Burying someone who is still alive
Watching child pornography.

References

Aharoni, E., Antonenko, O., & Kiehl, K. A. (2011). Disparities in the moral intuitions of criminal offenders: The role of psychopathy. *Journal of research in personality*, *45*(3), 322-327.

Axelrod, R. (2012). Launching 'the evolution of cooperation.' *Journal of Theoretical Biology, 299*, 21-24. doi: 10.1016/j.jtbi2011.04.015

Botvinick, M., Braver, T., Barch, D., Carter, C., & Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624-652.

Brainard, D. H. (1997). The Psychophysics Toolbox, *Spatial Vision, 10*, 433-436.

Buckholtz, 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*, 930-940. doi:10.1016/j.neuron.2008.10.016

Cannon, P., Schnall, S., & White, M. (2011). Transgressions and expressions: affective facial muscle activity predicts moral judgments. *Social Psychological and Personality Science 2*(3), 325-331. doi: 10.1177/1948550610390525

Cavanna, A., & Trimble, M. (2006). The precuneus: a review of its functional anatomy and behavioral correlates. *Brain, 129*, 564-583.

Chapman, H., Kim, D., Susskind, J., & Anderson A. (2009). In bad taste: evidence for the oral origins of moral disgust. *Science, 323*(5918), 1222-1226. doi: 10.1126/science.1165565

Cikara, M., Bruneau, E., & Saxe, R. (2011). Us and them: failures of intergroup empathy. *Current Directions in Psychological Science, 20*(3), 149-153. doi: 10.1177/0963721411408713

Chakroff, A. (2014) *Mapping morality from the ground up.* Unpublished manuscript.

Chakroff, A., & Young, L. L. (2014). *Harmful situations, impure people: an attribution asymmetry across moral domains.* Manuscript submitted for publication.

Conolly, A., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y., Wu, Y., Abdi, H., & Haxby, J. (2012). The representation of biological class in the human brain. *Journal of Neuroscience, 32*(8), 2608-2618. doi:10.1523/jneurosci.5547-11.2012

Contreras, J., Banaji, M., & Mitchell, J. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive Affective*

Neuroscience, 7(7), 764-770. doi: 10.1093/scan/nsr053

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* . (pp. 584 – 627). Hoboken, NJ: John Wiley & Sons.

Croft, K. (2009). *Exploring the role of ventromedial prefrontal cortex in human social learning: a lesion study.* (Doctoral dissertation). Retrieved from http://ir.uiowa.edu/etd/350

Cunningham, W., Raye, C., & Johnson, M. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience, 16*(10), 1717-1729.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion, 12*(1), 2-7. doi: 10.1037/a0025071

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

De Quervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305*(5688), 1254-1258.

FeldmanHall, O., Dagleish, T., Thompson, R., Evans, D., Schweizer, S., & Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive Affective Neuroscience, 7*(7), 743-751. doi: 10.1093/scan/nss069

Fellows, L., & Farah, M. (2007). The role of ventromedial prefrontal cortex in decision making: judgment under uncertainty or judgment per se? *Cerebral Cortex, 17*(11), 2669-2674. doi:10.1093/cercor/bhl176

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, *2*(4), 189-210.

Glenn, A. L., Iyer, R., Graham, J., Koleva, S., & Haidt, J. (2009). Are all types of morality compromised in psychopathy?. *Journal of personality disorders*, *23*(4), 384-398.

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5) 1029-1046. doi: 10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R. Koleva, S., & Ditto, P. H. (2011). Mapping

the moral domain. *Journal of Personal Social Psychology, 101*(2), 366-385. doi: 10.1037/a0021847

Greene, J. (2002). *The terrible, horrible, no good, very bad truth about morality and what to do about it*. (Doctoral dissertation). Retrieved from http://scholar.harvard.edu/joshuagreene/files/dissertation.pdf

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London Biological Sciences, 359*, 1775-1785. doi: 10.1098/rstb.2004.1546

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

Greene, J. (2007) The secret joke of Kant's soul. In Sinott-Armstrong, W. (Ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development.* Cambridge, MA: MIT Press, 35-117.

Haidt, J., & Graham, J. (2007) When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98.

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

Haidt, J. (2007). The new synthesis in moral psychology. Science, *316*, 998-1001. doi: 10.1126/science.1137651

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

Haidt, J. (2012). *The Righteous Mind*. New York: Vintage Books.

Hamann, K., Warneken, F., Greenberg, J., & Tomasello, M. (2011). *Nature, 476*(7360), 328-331. doi: 10.1038/nature10278

Hursthouse, R. (2012). Virtue Ethics. In Zalta, E. (Ed.), *The Stanford Encyclopedia of Philosophy.* Retrieved from http://plato.stanford.edu/archives/sum2012/entries/ethics-virtue/

Kanai, R., Feilden, T., Firth, C., & Rees, G. (2011). Political orientations are correlated with brain structure in young adults. *Current Biology, 21*, 677-680. doi:10.1016/j.cub.2011.03.017

Kemp, C., & Tenenbaum, J. (2008). The discovery of structural form. *PNAS, 105*(31), 10687-10692. doi:10.1073/pnas.0802631105

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446*(19), 908-911. doi:10.1038/nature05631

Kriegskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *PNAS, 103*(10), 3863-3868. doi:10.1073/pnas.0600244103

Kriegskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 1-28.

Lashkari, D., Sridharan, R., & Golland, P. (2010). Categories and functional units: an infinite hierarchical model for brain activations. *Advances in Neural Information Processing Systems, 23*, 1252-1260.

Lashkari, D., Vul, E., Kanwisher, N., & Golland, P. (2010). Discovering structure in the space of fMRI selectivity profiles. *NeuroImage, 50*, 1085-1098. doi:10.1016/j.neuroimage.2009.12.106

Lewis, G., Kanai, R., Bates, T., & Rees, G. (2012). Moral values are associated with individual differences in regional brain volume. *Journal of Cognitive Neuroscience, 24*(8), 1657-1663.

MacDonald, A., Cohen, J., Stenger, V. A., & Carter, S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(5472), 1835-1838.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143-152. doi: 10.1016/j.tics.2006.12.007

Mitchell, J., Banaji, M., & Macrae, C. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306-1315.

Moll, J., de Oliveira-Souza, R., Moll, F.T., Ignacio, F.A., Bramati, I., Caparelli-Daquer, E., & Eslinger, P. (2005). The moral affiliations of disgust: a functional MRI study. *Cognitive Behavioral Neurology, 18*(1), 68-78.

Moral Foundations. (n.d.). In MoralFoundations.org. Retrieved from www.moralfoundations.org

Moran, J., Young, L., Saxe, R., Lee, S.M., O'Young, D., Mavros, P., & Gabrieli J. (2010). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences, 108*(7), 2688-2692. doi: 10.1073/pnas.1011734108

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews, 9*, 856-869.

Nathaniel-James, D., & Frith, C. (2001). The role of the dorsolateral prefrontal cortex: evidence from the effects of contextual constraint in a sentence completion task. *NeuroImage, 16*, 1094-1102. doi: 10.1006/nimg.2002.1167

Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences, 8*(3), 102-107.

Olson, I., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain, 130*, 1718-1731. doi:10.1093/brain/awm052

Parkinson, C., Sinnott-Armstrong, W. Koralus, 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), 3162-3180. doi: 10.1162/jocn_a_00017

Paulus, M., & Frank, L. (2003). Ventromedial prefrontal cortex activation is critical for preference judments. *Brain Imaging, 14*(10), 1311-1315. doi: 10.1097/01.wnr.0000078543.07662.02

Pearson, J., Heilbronner, S., Barack, D., Hayden, B., & Platt, M. (2011). Posterior cingulate cortex: adapting behavior to a changing world. *Trends in Cognitive Sciences, 15*(4), 143-151.

Mansell, W. (2011). Control of perception should be operationalized as a fundamental property of the nervous system. *Topics in Cognitive Science, 3*, 257-261.

Rand, D., & Nowak, M. (2013). Human cooperation. *Trends in Cognitive Science, 17*(8), 413-425.

Redcay, E. (2008). The superior temporal sulcus performs a common function for social and speech perceptions: implications for the emergence of autism. *Neuroscience and Behavioral Reviews, 32*, 123-142.

Rilling, J., Sanfey, A. Aronson, J., Nystrom, L., & Cohen, J. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage, 22*, 1694-1703. doi: 10.1016/j.neuroimage.2004.04.015

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574-586.

Russell, P. S., & Giner-Sorolla, R. (2011). Social justifications for moral emotions: When reasons for disgust are less elaborated than for anger. *Emotion, 11*(3), 637.

Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. In Leslie, A., & German, T. (Eds.), *Handbook of Theory of Mind.* London, England: Psychology Press.

Shweder, R. A., Much, N.C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In Brandt, A., & Rozin, P. (Eds.) *Morality and Health*. New York: Routledge, 118-169.

Singh-Curry, V., & Husain, M. (2009). The functional role of the inferior parietal lobe in the dorsal and ventral stream dichotomy. *Neuropsychologia, 47*, 1434-1448. doi: 10.1016/j.neuropsychologia.2008.11.033

Smith, K., Oxley, D., Hibbing, M., Alford, J., & Hibbing, J. (2011). Disgust sensitivity and the neurophysiology of left-right political orientations. *PLoS ONE, 6*(11), 1-9.

Stanley, D., Sokol-Hessner, P., Fareri, D., Perino, M., Delgado, M., Banaji, M., & Phelps, E. (2012). Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society of London Biological Sciences, 367*, 744-753. doi: 10.1098/rstb.2011.0300

Tooby, J., & Cosmides, L. (2010). Groups in mind: the coalitional roots of war and morality. In Høgh-Olesen, H. (Ed.) *Human Morality and Sociality: Evolutionary & Comparative Perspectives*. New York: Palgrave MacMillan, 91-234.

Turiel, E. (1979). Distinct conceptual and developmental domains: Social convention and morality. In Howe, H., and Keasey, C. (Eds.), *Nebraska Symposium on Motivation, 1977: Social Cognitive Development*. Lincoln, NE: University of Nebraska Press, 77-116.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping, 30*, 829-858. doi: 10.1002/hbm.20547

Waytz, A., Zaki, J., & Mitchell, J. (2012). Response of dorsomedial prefrontal cortex predicts altruistic behavior. *The Journal of Neuroscience, 32*(22), 7646-7650.

Whitfield-Gabrieli, S., Moran, J., Nieto-Castanon, A., Triantafyllou, C., Saxe, R., & Gabrieli, J. (2011). Associations and dissociations between default and self-reference networks in the human brain. *NeuroImage, 55*, 225-232.

Wright, J. C., & Baril, G. (2011). The role of cognitive resources in determining our moral intuitions: Are we all liberals at heart?. *Journal of Experimental Social Psychology*, *47*(5), 1007-1012.

Wright, R. & Greene, J. (2013). *The Wright Show on Bloggingheads.tv, Oct 23, 2013 episode*. [Video] Retrieved from http://bloggingheads.tv/videos/22741

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgement. *NeuroImage 40*(4), 1912-1920. doi: 10.1016/j.neuroimage.2008.01.057

Young, L., Camprodon, J., 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 judgements. *Proceedings of the National Academy of Sciences, 107*(15), 6753-6758. doi:10.1073/pnas.0914826107

Zamboni, G., Gozzi, M., Krueger, F., Duhamel, J., Sirigu, A., & Grafman, J. (2012). Individualism, conservatism, and radicalism as criteria for processing political beliefs. *Social Neuroscience, 4*(5), 367-383.