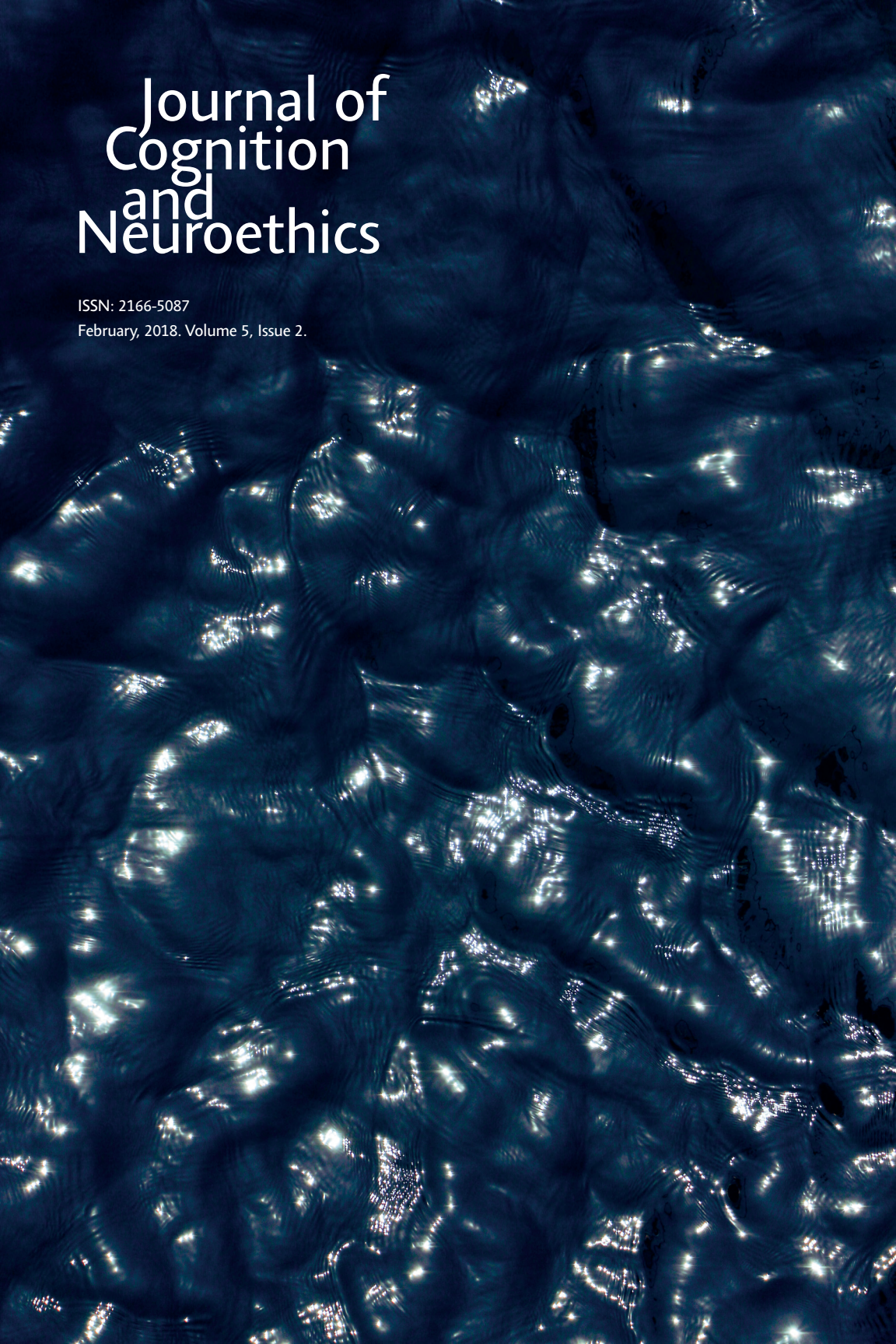
Journal of Cognition and Neuroethics

ISSN: 2166-5087
February, 2018. Volume 5, Issue 2.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 5, Issue 2 was digitally published in February of 2018 from Flint, Michigan, under ISSN 2166-5087.

The *Journal of Cognition and Neuroethics* is produced by the Center for Cognition and Neuroethics. For more on CCN or this journal, please visit cognethic.org.

Center for Cognition and Neuroethics
University of Michigan-Flint
Philosophy Department
544 French Hall
303 East Kearsley Street
Flint, MI 48502-1950

Table of Contents

Journal of Cognition and Neuroethics

Moral Mentation: What Neurocognitive Studies of Psychopathy May Really Offer the Internalism/Externalism Debate

Katherine L. Cahn-Fuller
Columbia University Medical Center

John R. Shook
University of Buffalo

James Giordano
Georgetown University

Acknowledgments
This work was supported in part by funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement 720270: HBP SGA1 (J.G.), and by federal funds UL1TR001409 from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through the Clinical and Translational Science Awards Program (CTSA), a trademark of the Department of Health and Human Services, part of the Roadmap Initiative, "Re-Engineering the Clinical Research Enterprise" (JG).

Biographies
Dr. Cahn-Fuller is a Resident in Psychiatry at Columbia University Medical Center, NY. Dr. Shook is research associate in philosophy and instructor of science education at the University of Buffalo, Buffalo, New York. Dr. Giordano is Professor in the Departments of Neurology and Biochemistry, and Chief of the Neuroethics Studies Program of the Pellegrino Center for Clinical Bioethics at the Georgetown University Medical Center, Washington, DC, and is a Research Fellow of the European Union Human Brain Project.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 2.

Citation
Cahn-Fuller, Katherine L., John R. Shook, and James Giordano. 2018. "Moral Mentation: What Neurocognitive Studies of Psychopathy May Really Offer the Internalism/Externalism Debate." *Journal of Cognition and Neuroethics* 5 (2): 1–20.

Moral Mentation: What Neurocognitive Studies of Psychopathy May Really Offer the Internalism/Externalism Debate

Katherine L. Cahn-Fuller, John R. Shook, and James Giordano

Abstract

A persistent debate about moral capacity – and neuroethics – focuses upon the internalism-externalism controversy. Internalism holds that moral judgments necessarily motivate an agent's actions; externalism views moral judgments as not inherently motivating an agent to perform moral actions. Neuroethical discussions of the putative cognitive basis of moral thought and action would be better informed if neurocognitive research would yield data sufficient for validating one side or the other. Neuroscientific studies of psychopaths have been employed in this regard. However, it seems that neuroscientific investigations to date have been inadequate to wholly define the nature of moral knowledge, and thus fail to preferentially support (or foster) an exclusively internalist or externalist view. Thus, moving forward it will be necessary to carefully define questions that neuroscience is employed to address and answer, and to ensure that empirical findings are not distorted to support preconceived theoretical assumptions. In this way, neuroscientific investigations can be used in a conciliatory way to both balance views of processes operative in moral cognition, and raise ethical, legal, and social questions about what research findings actually mean, and what medicine – and societies – will do with such information and meanings.

Keywords

Internalism, Externalism, Psychopathy, Neuroethics, Morality, Cognition, Emotion, Neuroscience

Introduction

Philosophers, legal scholars, social scientists, and psychologists have long questioned the nature of morality and the factors that drive moral judgment and behavior. The rise of biological psychology throughout the twentieth century provided a new lens through which to consider these questions and examine moral theories. Most recently, such pursuits have engaged the neurosciences in an attempt to develop empirically informed theories of moral cognition and action. Such neuroscientific studies of putative mechanisms of moral cognition and behavior has come to be regarded as one of the disciplinary foci of the field of neuroethics.

To be sure, the iterative use of neurotechnologies such as functional magnetic imaging (fMRI), forms of encephalography (namely quantitative electroencephalography

and magnetoencephalography), and neurogenetics have been instrumental to studies of the putative neural correlates of moral thought and behavior. In response to growing opportunities to experimentally formulate and test theories about moral cognition, neuroethics entails another – and perhaps more important – focus, namely, the analysis and proper interpretation and use of neuroscientific methods and data, inclusive of information about the possible neurocognitive processes involved in moral thought and action (Shook and Giordano 2015).

We do not share in the stern (neuro)skepticism that denies that neurotechnologically-enabled insights into brain architectures and processes should affect an understanding of morality. Morality is normative, the skeptics remind us, while scientific knowledge is merely factual – what we think is moral should not be warped by what we know is occurring. Skeptics have a logical point, but only that logical point. If moral ends are treated merely as ends, as static states of affairs either satisfied or not, then they do seem aloof from crude considerations about reaching them. What ought to be done cannot follow from what happens to be, if the skeptical point needs a logical axiom. Yet, that axiom is ambiguous, for there is a sense in which what ought to be can only follow from what happens – an outcome that ought to be done can only follow from preceding events that cumulatively produce that envisioned outcome. Expecting a future state of affairs to come about in the absence of a prior sequence of concrete matters producing that state of affairs is nothing short of expecting a miracle. No set of facts (i.e., what "is") logically implies what should morally happen (i.e., what "ought" to be). All the same, fulfilling what should morally happen surely implies some sequence of factual conditions yielding that outcome. An *ought* materially implies some *is*.

Due to that material implication, an adequate conception of a moral end includes some understanding of effective capacities for fulfilling it. Thinking about people attaining a moral end without any thought to relevant capacities that enable people to attain those ends is a diversion of mere imagination. Disassociating peoples' actual capacities from moral ends and then wondering why people's behaviors are more or less moral is even more futile. A conception of a capacity that has gone unrevised by available knowledge about brain activity controlling behavior is just an item of folk psychology: part of an interesting story about human nature that people have been in the habit of telling each other. Much of what has passed for moral philosophy, something too precious to compare with facts for those neuroskeptics, is nothing but idealized folk psychology. Moral philosophizing tends to fixate on ideal moral ends to the neglect of actual human capacities. Neuroethics need not be controlled by moral theories uninformed by

developmental psychology, cognitive psychology, and neuroscience (Shook and Giordano 2014).

The neuroskeptical worry that ethics will be dominated by neurology need not be indulged, either. There is plenty of middle ground for ethics and neuroscience to enhance each other. Ordinary moral psychology is vitally useful for human sociality, without question. Although hominid morality is far older than *Homo sapiens'* ability to talk about people being moral (or not moral), the proclivity for discussing how to be moral is not incidental. Labeling mental events accompanying behavioral habits intensifies their familiarity and allows better anticipatory control. Favoring or disfavoring recognizable mental outlooks thereby guides consequent behaviors; so effective mental outlooks are treated as "moral" insofar as they serve as capacities conducive to moral conduct. Furthermore, the only way to define a moral capacity in precise terms is to relate that capacity to a specific moral end. As a corollary, the contribution of a brain region's activity to a moral capacity cannot be estimated unless that moral capacity has been specified. Nothing happening in the brain correlates with "being nicer to people," but that is not because neuroscience failed to find some neuroanatomical locus for niceness. Rather, it is because "being nice" is just a vague behavioral end, not to mention an indistinct mental notion. Only a specific moral capacity for a concrete moral behavior can be experimentally associated with identifiable brain processes. This is not ontological reductionism, but only scientific empiricism.

Neuroethics can consult both neuroscience and moral psychology with growing confidence as their relevance to ethics becomes more interdependent. That interdependence is secured by their coordination with specific behavioral outcomes: well-defined moral ends. Neither neuroscience nor moral psychology by themselves should dictate matters to neuroethics. Detectable brain activities are not obviously *for* anything, and they mean very little until they are elicited through the exhibition of certain behavioral capacities; while competing views of psychological capacities that float freely from specified behavioral ends lead to abstract debating about what those affairs mean and what they are for. Specified links between psychological capacities and behavioral capacities are therefore essential for empirically associating brain processes with a person's ability to fulfill envisioned ends (Shook and Giordano 2016).

For example, there surely are many brain processes that allow one to fulfill a moral duty towards another. Learning which detectable neural activities in particular are essential processes needed for moral conduct is a goal that requires additional information. Certain brain processes do permit one's capacity for a specific kind of behavior, but it is not yet known whether such behavior is moral – is that behavioral

capacity deployed in order to be moral? (Conduct not done to be moral cannot be moral conduct even if there are morally good results.) And certain mental states do enable one's capacity to perform a specific moral end, but it's not yet known whether such states are moral – is that psychological capacity enacted due to morally worthy states? (Minding what one is doing to fulfill a moral end does not ensure that one is of a moral mind.)

Neuroethics should scrutinize claims about "moral" mental events that are allegedly due to brain processes "for" morality. Consider the following inference, and suppose that premises 1, 2, and 3 are all accurate:

1. Jo has the belief that s/he should state that X is morally right.
2. Jo does state that "X is morally right."
3. Brain activations A, B, C (etc.) are correlated with Jo's pronouncement that X is right.
4. A, B, and C are correlated with Jo's moral belief.
5. Therefore, A, B, and C are processes for morality.

Proposition 4 does not follow from 1–3, because "moral belief" is left ambiguous: is Jo's belief only "moral" because s/he says what we want her/him to rightly say, or because s/he sincerely wants to say what s/he personally thinks is morally right? We easily fill that gap with a fond intuition that 4 normally follows from 1–3. Yet, no given premise establishes a relationship between her/his psychological capacity and behavioral capacity in this situation. Hence, nothing can be concluded about the contributions made by A, B, and C to moral conduct. Experimental protocols should avoid these sorts of fallacious steps, and neuroethics should abstain from heedless theories about moral cognition.

To draw reasonable conclusions about the "moral" work of neurological processes and brain areas, an empirical investigation into the pursuit and achievement of moral ends requires some presumed information about a relationship – whether necessary or contingent, essential, or accidental – between a person's psychological and behavioral capacities. Where can that information come from? Folk psychology provides intuitive assumptions about the normal thinking and typical conduct of ordinary people in daily life. Has behavioral psychology better illuminated the reliable ties between what people affirm as moral and how they really behave themselves? Can neuroscience step in to expose the true connections between how brain works and what actions people perform? All three sources may be needed, although each has its weaknesses. Our neuroethical

concern here is whether recent claims made about moral neuroscience's contributions are fully warranted.

Tentative Roles of Emotion in Moral Decision-Making

Empirical studies examining relations between psychological affairs and behavioral capacities have shown how prosocial emotions, such as guilt and empathy, influence moral judgments and foster social cooperation and cohesion (Mendez 2009). Such research is of evident interest to social science, law, and politics. It may also elucidate ways that neural mechanisms are involved in abnormal psychological states associated with patterns of amoral and anti-social conduct. Prototypic in this regard are investigations of psychopathy that demonstrate psychopaths' atypical patterns of cognitive processing, as well as suggest that the anti-social behaviors characteristic of this condition result from insufficient affective/valuational input to moral decision-making processes (Harenski et al. 2010). Taken together, studies of psychopathy that employ neurotechnological methods may afford a better perspective on those ways that neuroscience could answer questions about the cognitive processes involved in (what is construed to represent) normal moral capacity and moral performance.

Findings that emotions guide moral judgments are hardly surprising given the belief that evolution has promoted cooperation and group solidarity through the development of prosocial dispositions (Mendez 2009). However, evidence of the extent of this connection prompts the question of whether emotions are a necessary feature of moral cognition. More specifically, it is important to ask if the absence of prosocial emotions or emotional dysfunction precludes the capability for moral knowledge. This question for moral neuroscience has practical implications for forensic psychiatry, as a patient's ability to understand moral norms may impact both their clinical and legal treatment. Information about the relationship between emotions and moral judgments can also influence views of, and beliefs about, moral behavior, theories of justice and normative theories of ethics. The role and relative importance of emotions to behavior are factors arousing intense debate.

One of the persistent debates at the core of theorizing about moral capacity revolves around the internalism-externalism controversy. In brief, internalism holds that moral judgments necessarily motivate an agent's actions, while externalism is the contrary view that moral judgments do not inherently motivate an agent to perform moral actions. Nuanced definitions for internalism and externalism, which receive some attention in this essay, incorporate refinements to anticipate objections. Neither side is immune to

empirical problems. For example, psychopathy appears to pose a serious problem for internalism, as the psychopath seems to make normal moral judgments but reveals and exhibits no motivation to act according to moral norms (Zhong 2013). In response to this problem, proponents of internalism have turned to neuroscience to demonstrate that the psychopath is incapable of forming genuine moral judgments due to emotional dysfunction. Proponents of externalism likewise employ empirical neuroscientific data for theoretical support, arguing that emotional input, while usually present, is not a necessary feature of moral knowledge.

If cognitive neuroscientific research would eventually yield the data sufficient for crediting one side or the other with better empirical warrant, then neuroethics should be most cautious among all the interested disciplinary viewpoints about offering any summary judgments. However, we think that some neuroethical perspective should be taken upon the way that the accumulation of studies to date is capable of supporting both internalism and externalism. Reflections on this empirical situation can go deeper than just noting the shifting tides to this debate. Questioning the interpretative assumptions made by both sides, and pondering the adoption of assumptions for framing empirical studies, led us to concur with the philosophical view that neuroscience alone cannot elucidate the nature of moral judgment and moral knowledge to the degree necessary for an adjudication of the internalism-externalism debate.

We begin with an overview of internalism and externalism, to set the stage for discussions of neuroscientific studies of psychopathy and their contributions to an understanding of moral decision-making. We then compare arguments for internalism and externalism, noting their common assumptions and divergent conceptual frameworks. Neuroscience by itself, we argue, will not dictate those methodological and interpretive terms, so neuroscience alone could never yield definitive support for either side of this internalist-externalist debate. There are no "theory-free" empirical results untouched by moral philosophizing here, whether from folk morality or academic ethics. Accordingly, we propose that the full incorporation of relevant empirical research calls into question whether a unitary conception of "moral judgment" has been in place all along.

Denying that neuroscience can decide a debate like internalism vs. externalism is not the larger moral of this story. Rather, neuroethical scrutiny such as ours can open opportunities for creatively re-framing what it actually means to form and act on moral judgments. If theoretical debates do persist, at least the different practical meanings attached to key terms such as "moral judgment" can be exposed and contrasted with clarity.

Studies of Psychopathy to Address the Internalism/Externalism Debate

The question of whether moral judgments are *necessarily* motivational has long been a contested issue in both moral philosophy and moral psychology. When an agent makes a sincere moral judgment, is she always motivated to abide by that judgment? Adina Roskies details this internalist thesis as follows: "If an agent believes that it is right to Φ in circumstances *C*, then s/he is motivated to Φ in *C*" (Roskies 2003). This does not mean, however, that all agents act to fulfill their moral judgments, as the motivation to do so may be outweighed by non-moral considerations. Rather, internalism simply requires that moral judgments elicit a *pro tanto* motivation that has some degree of compelling force. On the other hand, the thesis of motivational externalism, or "externalism," denies that the connection between moral judgments and motivation exists as a matter of conceptual necessity. The externalist position claims that if an agent is to be motivated to act in accordance with her moral judgment, then she must possess an additional desire that is external to the moral judgment itself. For example, the desire to do what is "good/right" can serve to motivate an agent to act according to her moral judgments (Rosati 2016).

Longstanding philosophical discourse has focused on the topic of moral motivation and approached the issue as a question to be answered through traditional reason and reflection. In recent years, philosophers have increasingly turned to empirical neuroscientific evidence to advance the discourse and attempt to resolve the debate. Scientific investigations of psychopathy have proven to be of keen interest in this regard, as the psychopath provides a real-world example of what amounts to an amoral agent. Psychopathy is characterized by the early onset of emotional, interpersonal, and behavioral dysfunctions, exemplified by lack of empathy and guilt, superficiality, unresponsiveness to relationships, grandiosity, and impulsivity (Cleckley 1988).

Recent neuroimaging studies have demonstrated a number of structural and functional abnormalities in brains of individuals with psychopathy, and the internalism/externalism debate has engaged such studies, among others, with considerable interest and enthusiasm. In what follows, we examine data from a number of neuroscientific studies examining psychopathy and reveal that empirical findings provide both support and contradictory evidence for internalism and externalism alike.

Support for Internalism

At the outset, we have reason to doubt that neuroscientific information can conclusively demonstrate that the link between moral judgment and motivation exists necessarily. As Jones has noted, showing an actual connection between judgment and

motivation does not show that it is necessary and holds in all possible worlds (Jones 2006). In this sense, empirical work in the neurocognitive sciences may be more relevant to externalism, as it need only demonstrate that the uncoupling of judgment and motivation is actual to show that it is possible. Internalists nevertheless utilize neuroscience to dissipate the threat posed the psychopath – the prototypical amoral agent who, despite possessing knowledge about moral norms, is not motivated to behave morally. To preserve the theory, philosophers have cited a number of neurocognitive studies as evidence that psychopaths do not make genuine moral judgments (Prinz 2006; Levy 2007), thereby preserving the essential link between judgment and motivation. A number of these studies have compared the neural networks engaged during moral decision-making in psychopathic and non-psychopathic populations.

Research has shown that moral judgments made by normal subjects are characteristically co-incident with emotion, suggesting that the motivational force of judgments is contingent on, or at minimum influenced by, emotions. Prinz has reviewed a number of neuroimaging studies measuring brain activity during morally neutral and morally valanced events and concluded that brain areas associated with emotional response were active when participants made moral judgments (Prinz 2006). The networks involving the amygdala are of interest in these studies, as such networks have been shown to be operative in the processing of emotional information (Blair 2005; Stratton, Kiehl, and Hanlon 2015). Further studies suggest that emotions not only co-occur but also influence the content of moral judgments. Schnall and colleagues demonstrated that the presence of an unpleasant odor or filthy surroundings made subjects more likely to condemn the actions described in a series of vignettes (Prinz 2006; Schnall et al. 2008; Sauer 2012). Wheatley and Haidt showed that experimentally augmented feelings of disgust altered subjects' moral judgments (Prinz 2006; Sauer 2012; Wheatley and Haidt 2005). The results of these studies support the hypothesis that emotions influence and may increase the severity of one's moral judgments. So while an exact role of emotions in moral decision-making remains unclear – and a matter of debate (Zhong 2013; Prinz 2006; Sauer 2012) – there is growing agreement that emotions play a critical role in the formation of moral judgments, at least in normal individuals.

However, the emotional input that is characteristic of moral judgments in normal individuals tends to be markedly absent in the psychopath (Harenski et al. 2010; Blair 2005; Stratton, Kiehl, and Hanlon 2015). There is considerable literature to suggest that cardinal traits of psychopathy (e.g., lack of empathy, remorse, guilt, and shallow affect) reflect, and/or are the product of, dysfunction of networks involving the amygdala. Structural imaging studies reveal that individuals with robust psychopathic traits have

decreased volume and morphological deficits of the amygdala (Stratton, Kiehl, and Hanlon 2015). Functional neuroimaging investigations have demonstrated reduced amygdalar activation during the processing of affective stimuli when adult psychopaths are asked to rate the severity of moral violations (Harenski et al. 2010, Stratton, Kiehl, and Hanlon 2015). Psychopaths also show impairment on aversive conditioning and passive avoidance learning tasks, both of which are reliant upon functional integrity of amygdala networks (Blair 2005). In addition to amygdalar dysfunction, psychopaths also display reduced activation of the ventromedial prefrontal cortex (vmPFC) in response to emotional stimuli. Given efferent connections between the amygdala and the vmPFC, Blair hypothesized that moral attitudes may be reliant upon stimulus-outcome processing subserved by an amygdalar-vmPFC network (Stratton, Kiehl, and Hanlon 2015, Blair 2008). In this model, amygdalar activation by a conditioned stimulus provides input to the vmPFC, which represents this information as a valenced outcome. This process, thought to be essential for moral reasoning, is dysfunctional in psychopaths.

In that the characteristic features of psychopathy are due, in part, to severe emotional dysfunction, and because emotions play a critical role in moral judgments in normal subjects, some philosophers have turned to psychopathy literature to support internalism. Prinz argues that the psychopath's emotional deficiencies prevent him from making genuine moral judgments (Prinz 2006). To support this conclusion, he points to Blair's studies demonstrating that psychopaths have difficulty recognizing negative emotions in others, are not amenable to fear conditioning, experience pain less intensely than normal subjects, and are undisturbed by distressing photographs (Blair et al. 1997; Blair et al. 2001; Blair et al. 2002). Unable to experience fear, empathy, remorse, and guilt, the psychopath lacks the moral knowledge required to make genuine moral judgments. While the psychopath may acknowledge that certain criminal acts are 'wrong,' Prinz denies that such moral statements constitute genuine beliefs in the absence of emotions, stating: "Can one sincerely attest that killing is morally wrong without being disposed to have negative emotions towards killing? My intuition here is that such a person would be confused or insincere" (Prinz 2006, 32). The claim fortifies an internalist stance, in that, it argues that psychopaths are unmotivated by moral norms because they are incapable of forming genuine moral judgments.

An appeal to the intuition that psychopaths do not make genuine moral judgments as evidence for the necessity of emotions for moral judgments begs the question. This is clear if the argument is distilled as follows:

P1: Psychopaths have no negative emotions, such as fear

P2: Psychopaths do not make genuine moral judgments *(intuition)*

C: Emotions are necessary for moral judgments

Prinz advances the claim that psychopaths' emotional dysfunction precludes them from understanding morality. While neuroscientific studies demonstrate that psychopaths lack patterns of amygdalar-vmPFC activity involved in emotional aspects (and influence) of moral thought, neuroscience does not, and likely cannot, define what is required to make an 'authentic' moral judgment. There is thus insufficient neuroscientific evidence to derive the conclusion that the emotional abnormalities of the psychopath prevent the acquisition of moral knowledge. In response to Prinz's claims, an externalist could (and likely would) simply dispute his intuition (Liao 2016).

Another argument often used to defend internalism is based upon the inability of psychopaths to distinguish moral and conventional transgressions. Apropos, Blair assessed psychopaths' response to the moral/conventional transgression task (MCT), a test initially developed to determine if children could distinguish between moral and conventional transgressions (Blair 1995; Blair et al. 1995; Nucci and Turiel 1978; Shoemaker 2011). The test requires subjects to: (1) determine if the action in the scenario is permissible; (2) rate the seriousness of the transgression; (3) justify why an action was or was not permissible; and, (4) determine if the wrongness of the action is dependent on an authority figure. Results demonstrated that psychopaths, unlike non-psychopathic children and adults, judged moral and conventional transgressions similarly and were less likely to justify their judgments with reference to the victim's welfare. Interestingly, psychopaths judged all transgressions to be authority-independent, a quality usually assigned to only moral transgressions. This finding disproved Blair's prediction that psychopaths would declare both moral and conventional transgressions to be authority-dependent. Blair interpreted this tendency as the psychopaths' desire to demonstrate they had reformed and learned the rules of society, causing them to overcompensate and declare all transgressions were authority-independent rather than risk classifying moral transgressions as authority-dependent.

Levy has noted that psychopaths' performance on the MCT provides evidence that they lack moral knowledge, thereby endorsing the view that psychopaths are incapable of forming authentic moral judgments and supporting the internalist stance (Levy 2007). Levy fortifies these assertions with neuroscientific findings about dysfunction of amydalar networks in psychopaths, which contributes to their inability to categorize harms in terms of their effect on the emotional states of others. Because psychopaths are

unable to grasp the distinct nature of moral transgressions, Levy posits that they have a reduced sense of moral responsibility. Sauer also employed the MCT as a test of one's ability to form moral judgments, stating:

> Moral judgment requires the capacity to understand a certain subclass of prescriptive social rules as non-conventional, transgressions of these norms as more serious, generalizable wrong… and the validity of these rules as neither based on social acceptability nor dependent on authority. (Sauer 2012, 98)

Given these criteria, the ability to understand differences between moral and conventional transgressions becomes an essential element of moral judgment. Pro Levy, Sauer relates the psychopath's failure to distinguish moral and conventional transgressions to their lack of emotions and inability to perceive the "special" character of their violations. He concludes, pro the internalist view, that emotional responsiveness is necessary for moral judgment.

However, recent empirical work has questioned the validity of the MCT. Aharoni et al. used a modified version of the MCT to assess moral decision-making in 109 incarcerated psychopathic offenders (Aharoni, Sinnott-Armstrong, and Kiehl 2012; Godman and Jefferson 2017). This version of the MCT employed a forced-choice method in which subjects were informed that exactly half of the test scenarios were morally wrong, removing the incentive to over-rate all acts as moral transgressions. The authors found that performance on the task was not related to psychopathy scores, but was instead correlated with IQ and antisocial characteristics. Dolan and Fullam re-assessed the MCT in adolescent offenders and also failed to find an association between psychopathic traits and task performance (Dolan and Fullam 2010; Levy 2014).

Shoemaker has directly criticized the significance of the moral/conventional distinction, arguing that the distinction is not reflective of a unitary concept but represents, rather, several sub-distinctions that sometimes overlap (Shoemaker 2011; Godman and Jefferson 2017). He deconstructed the moral/conventional distinction into 3 primary dimensions: (1) the permissible/impermissible distinction; (2) the more serious/less serious distinction; and, (3) the authority dependent/authority independent distinction. These sub-distinctions do not necessarily map onto each other or the overall moral/conventional distinction, causing Shoemaker to conclude that the moral/conventional distinction cannot bear the weight of determining the moral responsibility of psychopaths.

Even if the moral/conventional distinction and evidence of psychopaths' inability to make this differentiation were upheld, we believe that these finding are insufficient to conclude that psychopaths lack moral knowledge. The inability to distinguish between moral and conventional transgressions is certainly significant to the internalism/externalism debate, insofar as it demonstrates a deficiency in the kind of moral understanding that is required to make moral judgments. However, we maintain that psychopaths' performance on the MCT does not provide adequate stand-alone evidence that they lack of moral knowledge.

Moral transgressions can be defined by their consequences for the rights and welfare of others, and are differentiated from conventional transgressions by the presence of a victim (Blair 1995). Given what is known about psychopaths' emotional (dys)function, it is unsurprising that studies have shown psychopaths to be significantly less likely to justify their judgments by reference to the victim's welfare. Such findings do not, however, indicate that psychopaths are *incapable* of identifying victims or the emotional state of others. Indeed, many psychopaths explained their judgments with reference to a victim's welfare (Blair 1995), and a number of studies demonstrate that psychopaths are capable of evaluating the emotions of others. Decety and colleagues presented visual depictions of social interactions to psychopaths and found that those with a high-level of psychopathy accurately identified the emotions of the subjects in the interactions, including the victims of harmful actions and the recipients of helpful actions (Decety et al. 2015). Dolan et al. found that psychopathic traits were not associated with marked difficulties in reading basic and complex emotions from facial expression (Dolan and Fullam 2004).

These empirical results suggest that, despite their decreased empathy, psychopaths possess knowledge of others' thoughts and feelings. So, while empathy and other emotions may be important (if not required) to motivate psychopaths to act according to their moral judgments, empirical findings do not support that these qualities are wholly necessary for moral judgment itself.

Support for Externalism

As previously stated, neuroscientific evidence may be somewhat more useful to support an externalist view, which needs only to show that separation between moral judgment and motivation is *possible*. Of note, this does not obligate the belief that judgment and motivation are not *usually* linked but that this link is not a conceptual necessity. Because psychopathic criminal offenders provide real-world examples of a lack

of moral motivation, externalists turn to neuroscientific information about processes involved in and/or subserving interactions of emotion, decision-making and actions as evidence that psychopaths form genuine moral judgments.

Evidence for this is provided by studies demonstrating that psychopaths make the same moral judgments as non-psychopathic individuals. For example, Glenn et al. presented psychopaths with personal moral dilemmas (defined as those involving salient harm to another individual), impersonal moral dilemmas (those not involve harm to another individual), and non-moral dilemmas (Glenn, Raine, and Schug 2009). While neuroimaging demonstrated that psychopaths had reduced amygdalar activity during emotional moral decision-making, there was no significant relationship between psychopathy scores and the proportion of utilitarian responses to personal moral dilemmas (Glenn et al. 2009). Cima et al. found similar results (Cima, Tonnaer, and Hauser 2010): psychopaths, like healthy subjects and non-psychopath delinquents, judged impersonal moral actions to be more permissible than personal moral actions, even though both types of harms led to utilitarian gains. Furthermore, there were no group differences in moral judgments for either impersonal or personal scenarios, with psychopaths no more likely to support utilitarian outcomes than other test subjects. Cima and colleagues concluded that these results do not support the hypothesis that emotional processes are necessary for moral judgments, but instead indicate that psychopaths understand distinctions between right and wrong but do not use such knowledge to guide their actions.

These findings suggest that psychopathic individuals use alternative strategies to compensate for their diminished emotional processing, enabling them to make moral judgments. Indeed, Glenn and colleagues found that psychopathy is associated with increased activity in the dorsolateral prefrontal cortex during moral decision-making (Glenn et al. 2009). Likewise, Kiehl et al. demonstrated increased activation of cortical regions in psychopaths during processing of affective stimuli (Kiehl et al. 2001). Such studies suggest that psychopaths rely heavily on abstract reasoning to process moral information. Glenn and colleagues summarized the findings of these cognitive and imaging studies:

> Although [psychopaths] may cognitively *know* the difference between right and wrong (i.e., the moral judgment), they may not have the *feeling* of what is right and wrong, and thus lack the motivation to translate their moral judgments into appropriate moral behavior. (Glenn et al. 2009, 910)

Cima et al. agreed, stating that normal emotional processing may be unnecessary for forming moral judgments, yet is likely important in generating an appreciation of moral distinctions and in guiding actions (Cima, Tonnaer, and Hauser 2010).

Such studies may offer evidence that psychopaths do in fact make genuine moral judgments, thus upholding both the psychopath as a paradigm for the separation of moral judgment and motivation and the externalist view. Yet, these studies, like those cited by the internalists, do not answer the (primarily conceptual) question: *What is a genuine moral judgment?* Assuming that the presented data are accurate, it becomes clear that psychopaths respond to moral dilemmas in a manner similar to non-psychopathic controls, despite differences in patterns of amygdalar and prefrontal cortical network activation.

Conclusions: Toward a Conciliatory View – and Approach

This empirical situation leads us to ask if the *source* of moral judgments is essential to their authenticity. Presumably, the internalist view would argue that moral judgments that result from abstract reasoning processes rather than emotional input are not 'authentic.' Of course, this statement, as we have seen, begs the question at hand. But the externalist view is mistaken to conclude that psychopaths possess true moral knowledge by virtue of the fact that they verbally respond to moral dilemmas in the same way as controls. This conclusion is grounded in the assumption that moral knowledge is not contingent on a particular thought process, which is the premise that internalists reject when they cite the emotional input that characterizes normal decision-making. Thus, it seems that neuroscientific investigations to date have been inadequate to wholly define the nature of moral knowledge, and therefore fail to preferentially support (or foster) an exclusively internalist or externalist view.

We have pointed out ways that neuroscientific evidence, by itself, does not appear to be sufficient for describing the nature of moral knowledge. This does not mean, however, that the internalism/externalism debate has nothing to gain from neuroscience. To the contrary, studies of the neural networks involved in moral cognition reveal two important findings. First, emotional input is a feature of moral judgments in non-psychopathic individuals. Second, the emotional dysfunction of psychopaths correlates with the absence of moral motivation. These data focus the debate and lead us to question if emotions, understood as one of many inputs to (moral) decision-making processes, are essential to the formation of authentic moral judgments. The link between (moral) emotions and compliance with moral norms is notably significant to psychiatry,

as it informs predictions about the relative validity and value of therapeutic interventions intended to mitigate or prevent psychopathic behaviors.

Neuroscientific studies, such as those discussed, can also call into question any strict determination of the interalism/externalism disagreement. For example, it may be the case that humans are not universally motivated, or unmotivated, by moral judgments. Rather, the degree to which moral judgments motivate agents to act may differ across circumstances and individuals. Zhong takes this approach, arguing that emotions, while not causally necessary for moral judgment, may titrate the severity of moral judgment (Zhong 2013). On this view, the emotions associated with a moral judgment will influence the extent to which the judgment overrides other considerations in favor of an action. To support his claims, Zhong points to studies demonstrating that psychopaths and non-psychopaths often make similar moral judgments, explaining these findings with reference to the cognitive, non-emotional mechanisms that both groups use to process moral information. Emotional input is therefore significant to moral motivation insofar as it alters the severity of moral judgments.

Even if we do not accept this argument, we have reason to question whether *all* moral judgments made by non-psychopaths are dependent on emotional input. Let us consider two ways that this might not be the case. First, there may be a subset of moral dilemmas that do not provoke a significant emotional response. Empirical data already support this claim. Studies by Greene and colleagues, for example, revealed that normal subjects' brain regions show similar patterns of activity when these subjects respond to impersonal moral dilemmas and non-moral dilemmas (Greene et al. 2001). Unless judgments elicited by impersonal moral dilemmas do not constitute authentic moral judgments, this finding gives us reason to doubt that emotional input is necessary for all moral understanding. Second, the emotions triggered by moral dilemmas may be morally irrelevant. On this reading, the presence of moral emotions should have no impact on an agent's moral judgments. Greene and Singer take this stance, arguing that moral emotions are an evolutionary byproduct and fail to track "moral truths" (Singer 2005; Greene 2008). As such, there may be reason to ignore emotionally driven moral intuitions in favor of more reasoned conclusions.

Continued investigations of brain structures and functions that are involved in moral cognition are sure to advance this discussion. The information gained from these studies is important not only to the philosophical debate at hand but also to forensic psychiatry and the justice system which look to empirical data about psychopathy to inform judgments about criminal responsibility and what could – and should – be done about criminal behavior (Giordano, Kulkarni, and Farwell 2014). It is important to remember,

however, that neuroscience is unlikely to provide definitive answers to the conceptual questions that drive the current version of the internalism/externalism debate. Moving forward, it will therefore be necessary to carefully define the questions that neuroscience is employed to address and answer, and equally vital to ensure that empirical findings are not distorted to support preconceived theoretical assumptions. In this way, neuroscientific investigations can be used in a conciliatory way. Not only to balance views of processes operative in moral cognition, but to bring together the sciences and humanities to both address questions about human morality, and iteratively raise ethical, legal and social questions about what research findings actually mean, and what medicine – and societies – will effect through the use of such information and meanings.

References

Aharoni, Eyal, Walter Sinnott-Armstrong, and Kent A. Kiehl. 2012. "Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction." *Journal of Abnormal Psychology* 121 (2): 484–497.

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

Blair, R. J. R. 2005. "Applying a cognitive neuroscience perspective to the disorder of psychopathy." *Development and Psychopathology* 17 (3): 865–891.

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

Blair, R.J.R., D.G.V. Mitchell, R.A. Richell, S. Kelly, A. Leonard, C. Newman, and S.K. Scott. 2002. "Turning a deaf ear to fear: Impaired recognition of vocal affect in psychopathic individuals." *Journal of Abnormal Psychology* 111 (4): 682–686.

Blair, R.J.R, E. Colledge, L. Murray, and D.G. Mitchell. 2001. "A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies." *Journal of Abnormal Child Psychology* 29 (6): 491–498.

Blair, R.J.R., L. Jones, F. Clark, and M. Smith. 1995. "Is the Psychopath 'morally insane'?" *Personality and Individual Differences* 19 (5): 741–752.

Blair, R.J.R., L. Jones, F. Clark, and M. Smith. 1997. "The psychopathic individual: A lack of responsiveness to distress cues?" *Psychophysiology* 34 (2): 192–198.

Cima, Maaike, Franca Tonnaer, and Marc D. Hauser. 2010. "Psychopaths know right from wrong but don't care." *Social Cognitive and Affective Neuroscience* 5 (1): 59–67.

Cleckley, Hervey Milton. (1941) 1988. *The Mask of Sanity*. 5th Edition. St. Louis, MO: Mosby.

Decety, Jean, Chenyi Chen, Carla L. Harenski, and Kent A. Kiehl. 2015. "Socioemotional processing of morally-laden behavior and their consequences on others in forensic psychopaths." *Human Brain Mapping* 36 (6): 2015–2026.

Dolan, Mairead C., and Rachael S. Fullam. 2004. "Theory of mind and mentalizing ability in antisocial personality disorders with and without psychopathy." *Psychological Medicine* 34 (6): 1093–1102.

Dolan, Mairead C., and Rachael S. Fullam. 2010. "Moral/conventional transgression distinction and psychopathy in conduct disordered adolescent offenders." *Personality and Individual Differences* 49 (8): 995–1000.

Giordano, James, Anvita Kulkarni, and James Farwell. 2014. "Deliver us from evil? The temptation, realities and neuroethico-legal issues of employing assessment neurotechnologies in public safety initiatives." *Theoretical Medicine and Bioethics* 35 (1): 73–89.

Glenn, A.L., A. Raine, and R.A. Schug. 2009. "The neural correlates of moral decision-making in psychopathy." *Molecular Psychiatry* 14 (1): 5–6.

Glenn, A.L., A. Raine, R.A Schug, L. Young, and M. Hauser. 2009. "Increased DLPFC activity during moral decision-making in psychopathy." *Molecular Psychiatry* 14 (10): 909–911.

Godman, Marion, and Anneli Jefferson. 2017. "On Blaming and Punishing Psychopaths." *Criminal Law and Philosophy* 11 (1): 127–142.

Greene, Joshua. 2008. "The Secret Joke of Kant's Soul." In *Moral Psychology*, edited by Walter Sinnott-Armstrong, 35–79. Cambridge, MA: MIT Press.

Greene, Joshua D., Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (5537): 2105–2108.

Harenski, Carla L., Keith A. Harenski, Matthew S. Shane, and Kent A. Kiehl. 2010. "Aberrant neural processing of moral violations in criminal psychopaths." *Journal of Abnormal Psychology* 119 (4): 863–874.

Jones, Karen. 2006. "Metaethics and emotions research: A response to Prinz." *Philosophical Explorations* 9 (1): 45–53.

Kiehl, Kent A., Andra M. Smith, Robert D. Hare, Adrianna Mendrek, Bruce B. Forster, Johann Brink, and Peter F. Liddle. 2001. "Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging." *Biological Psychiatry* 50 (9): 677–684.

Levy, Neil. 2007. "The Responsibility of the Psychopath Revisited." *Philosophy, Psychiatry, & Psychology* 14 (2): 129–138.

Levy, Neil. 2014. "Psychopaths and blame: The argument from content." *Philosophical Psychology* 27 (3): 351–367.

Liao, Matthew S. "Empirical Science and Motivation Internalism." Lecture, New York University, New York, NY, February 2016.

Mendez, Mario. 2009. "The Neurobiology of Moral Behavior: Review and Neuropsychiatric Implications." *CNS Spectrums* 14 (11): 608–620.

Nucci, Larry P., and Elliot Turiel. 1978. "Social Interactions and the Development of Social Concepts in Preschool Children." *Child Development* 49 (2): 400–407.

Prinz, Jesse. 2006. "The emotional basis of moral judgments." *Philosophical Explorations* 9 (1): 29–43.

Rosati, Connie S. "Moral Motivation." *Stanford Encyclopedia of Philosophy*. Published July 7, 2016. Accessed March 2017. https://plato.stanford.edu/entries/moral-motivation/.

Roskies, Adina. 2003. "Are ethical judgments intrinsically motivational? Lessons from 'acquired sociopathy'." *Philosophical Psychology* 16 (1): 51–66.

Sauer, Hanno. 2012. "Psychopaths and Filthy Desks." *Ethical Theory and Moral Practice* 15 (1): 95–115.

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2008. "Disgust as Embodied Moral Judgment." *Personality and Social Psychology Bulletin* 34 (8): 1096–1109.

Shoemaker, David W. 2011. "Psychopathy, Responsibility, and the Moral/Conventional Distinction." *The Southern Journal of Philosophy* 49 (s1): 99–124.

Shook, John R., and James Giordano. 2014. "A Principled and Cosmopolitan Neuroethics: Considerations for International Relevance." *Philosophy, Ethics, and Humanities in Medicine* 9 (1).

Shook, John R., and James Giordano. 2015. "Minding Brain Science in Medicine: On the Need for Neuroethical Engagement for Guidance of Neuroscience in Clinical Contexts." *Ethics in Biology, Engineering and Medicine* 6 (1–2): 37–42.

Shook, John R., and James Giordano. 2016. "A Neuropragmatic and Neuro-Ecological Approach to Neuroethics." *Pragmatism Today* 7 (1): 20–31.

Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9 (3/4): 331–352.

Stratton, John, Kent A. Kiehl, and Robert E. Hanlon. 2015. "The Neurobiology of Psychopathy." *Psychiatric Annals* 45 (4): 186–194.

Wheatley, Thalia, and Jonathan Haidt. 2005. "Hypnotic Disgust Makes Moral Judgments More Severe." *Psychological Science* 16 (10): 780–784.

Zhong, Lei. 2013. "Internalism, Emotionism, and the Psychopathy Challenge." *Philosophy, Psychiatry, & Psychology* 20 (4): 329–337.

Journal of Cognition and Neuroethics

A Neuroscience Study on the Implicit Subconscious Perceptions of Fairness and Islamic Law in Muslims Using the EEG N400 Event Related Potential

Ahmed Izzidien
The University of Cambridge

Srivas Chennu
The University of Kent
The University of Cambridge

Acknowledgments
We would like to express our gratitude to the Cambridge Muslim College for allowing us to undertake several of the studies on their premises. We are also grateful to Dr. Timothy Winter for suggesting the cognition of religion as a topic of research.

Biographies
Dr. Ahmed Izzidien is a visiting researcher at the Cambridge Forum for Legal and Political Philosophy, Faculty of Law, the University of Cambridge. His interests include legal hermeneutics, implicit values and the cognition of religion and law. Dr. Srivas Chennu is a lecturer at the University of Kent and a senior research associate at the University of Cambridge. His research focuses on how the brain mechanisms underlying consciousness are altered in sleep, sedation, meditation, and the vegetative and minimally conscious states.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 2.

Citation
Izzidien, Ahmed, and Srivas Chennu. 2018. "A Neuroscience Study on the Implicit Subconscious Perceptions of Fairness and Islamic Law in Muslims Using the EEG N400 Event Related Potential." *Journal of Cognition and Neuroethics* 5 (2): 21–50.

A Neuroscience Study on the Implicit Subconscious Perceptions of Fairness and Islamic Law in Muslims Using the EEG N400 Event Related Potential

Ahmed Izzidien and Srivas Chennu

Abstract

We sought to compare the implicit and explicit views of a group of Muslim graduates on the fairness of Islamic law. In this preliminary investigation, we used the Electroencephalographic N400 Event Related Potential to detect the participant's implicit beliefs. It was found that the majority of participants, eight out of ten, implicitly held that Islamic Law was unfair despite explicitly stating the opposite. In seeking to understand what separated these eight participants from the remaining two – the two who both implicitly and explicitly held that Islamic Law was fair – only two distinguishing characteristics could be identified. Both participants had undertaken an in-depth study of a branch of Islamic law that places the spirit of the law above that of a literal interpretation. They had also attended the same seminary, exclusive to the other participants. Of the eight participants, it was discovered that, while they implicitly held Islamic law to be unfair, they also held it to be rational – in the same way they found that it was rational to push a person off a ship in order to save the remaining from drowning, yet unfair. We discuss these preliminary findings and consider theories on how an innate sense of fairness, an aspect of nativism, may come into play when it is not congruent with a participant's own beliefs. Further, we ask, where such an inconsistency occurs, how does the mind attempt to rectify it – if at all? As a possible contribution to the discussion on theories of nativism vs. empiricism we put forward a hypothesis and methodology for investigation that may produce previously unconsidered data on human nature.

Keywords

Law, EEG, N400, Islam, Maqasid, Implicit, Values

Introduction

Children often question why things are the way they are. They also expect to be treated fairly amongst their peers. Such qualities, observed in infants from as young as six months, have spurred a theory that teleological reasoning and fairness are both innate (J. K. Hamlin 2015) (Deborah Kelemen and DiYanni 2005) challenging the empiricist notion that such qualities are learned. As infants progress through childhood and into adulthood, these qualities are seen to persist (Poling and Evans 2002) (Lombrozo, Kelemen, and Zaitchik 2007) (Deborah Kelemen, Rottman, and Seston 2013). It is such

that even with the most Machiavellian, seeing unfairness as justified for the gain of power, such unfairness is never wished upon themselves. An innate aversion to unfairness appears to persist.

Here we pose a question: what occurs implicitly within the mind of an individual who subscribes to an authority that defines its own version of fairness and purpose? For example, an individual may subscribe to a political ideology that sets policy that it then defines as fair and of service to a civic purpose. These policies may often be framed by the said authority as being beyond the rational grasp of the said individual and to be taken on trust despite any internal personal reservations. Another example may be that of a religious authority that defines its commands as fair and of purpose. Followers would be expected to accept such designations, even if they could not rationalise them. Thus, to consider how the mind of a person responds in this context, we sought to use the authority of 'Islam' defining its laws as 'fair and of purpose.' Using only Muslim participants for the study, the focus was on the authority of their religion defining the fairness and purpose of Islamic law. What their religion deemed to be fair and purposeful was to be accepted as fair and purposeful even if they could not rationalise such designations.

In order to measure the implicit response of the human mind towards such designations that are set by the authority to which they subscribe, we measured the implicit attitude of the individuals under investigation. This is because naturally occurring implicit attitudes have been found to be a more accurate measure of attitude than direct measures such as survey items with summated rating scales (Graham et al. 2012). Responses on direct measures such as surveys can represent conscious evaluations of content in memory rather than its activation (Gawronski, LeBel, and Peters 2007). Whereas many values, attitudes, and goals operate at implicit levels (Johnson and Saboe 2011) (Bargh and Chartrand 1999) (Greenwald and Banaji 1995), and often occur outside people's awareness, intention, and control (Wittenbrink and Schwarz 2007) (Johnson and Saboe 2011). Furthermore, there is increasing evidence that much affective and cognitive regulation occur automatically (Kanfer 2009), and while justice researchers often limit their attention to the explicit level, it is likely that justice has implicit effects because fairness-related experiences involve conditions of high arousal and strong affect (Tripp, Bies, and Aquino 2002) which increase the likelihood of implicit processing (Metcalfe and Mischel 1999) (Johnson and Saboe 2011).

Ordinarily, observant Muslims would explicitly respond that 'Islamic law' was fair. The source texts of the religion often remind of a critical importance of justice and rational purpose behind law. However, for this investigation, we were wholly unaware of

the implicit attitudes of Muslims in this area as no such study has as yet been undertaken, particularly not from the perspective of the cognition of law and religion. Discussions around fairness, purpose, and authority in approaching Islamic source texts have to date often been based in philosophical and theological discourse with little perspective from the natural sciences.

A new assessment on how particular approaches to understanding Islamic law manifest in the mind may lead to new perspectives on how the mind responds to commands that do not necessary align with one's innate disposition. It may also help to inform how emerging Muslim majority counties in the world address this issue on a legal and constitutional level, particularly as fairness has been seen as a factor in bringing people together (Johnson and Lord 2010).

Furthermore, while studies have documented how fairness judgments in general affect policy positions, there has been relatively little done on the genesis of the fairness judgments themselves. Lind has proposed that justice perceptions are pivotal cognitions because they prime motivations that give rise to specific behaviors (2002, 67). Studies have also found substantial support for proposed links between the implicit effects of justice and self identity (Cropanzano et al. 2001).

Thus, detecting the attitudes of individuals towards law – with a commitment to a specific authority – may shed some light on how determinations of fairness and purpose in law may be pre-set in some individuals even before they evaluate cases of law, as will be discussed.

To assess why participants may hold the implicit attitudes that are contrary to their explicit attitudes, we also collected information on their religious education background and how they normally approached legal problems such as the trolley problem using an anonymous questionnaire. Being unaware of the implicit responses of Muslim participants to questions on the fairness and purpose of Islamic law, this being the first study of its kind as mentioned, we approached the study inductively. Then, using the data from the outcome, we formed two hypothesis for further study, one contingent on the implicit vs. explicit findings.

To measure implicit attitudes, we would use the EEG N400 event related potential (ERP) method. The method allows us to discern if a person's expressed explicit attitude is the same as their implicit attitude (Lind 2002) (Van Berkum et al. 2009) (Leuthold et al. 2015). For example, a study (Van Berkum et al. 2009) asked two groups of participants to consider the statements: (a) "I think euthanasia is an acceptable course of action;" and, (b) "I think euthanasia is an unacceptable course of action." For the twenty-one respondents the study describes as the strict Christian (SC) group, the authors compared

the Event Related Potential (ERP) responses to the value-inconsistent word 'acceptable' in the first sentence with ERP responses to the value-consistent word 'unacceptable' in the second sentence. For the twenty-one respondents of the non-Christian (NC) group, they compared ERPs across the same statements. They found that value-inconsistent sentences increased the amplitude of the N400 component (Van Berkum et al. 2009) a finding also observed in a study that examined the N400 for value-inconsistent sentences (Leuthold et al. 2015). The N400 response is a broad negative signal that appears around 400ms after the test word has been presented aurally or visually (Lau, Phillips, and Poeppel 2008). The system used has been further investigated by Wiswede who found the N400 marker can only be obtained when a participant uses an evaluative mindset (Wiswede et al. 2013). Thus, for this study we measured the ERP of Muslim participants when they considered the sentences 'I believe Islamic law is fair/unfair.'

Materials and Methods

Participants

Ethical approval from the University of Cambridge Psychology Research Ethics Committee was obtained for the study. As the study required the participation of committed and practicing Muslims, we advertised in the local Muslim Society and at two mosques in the Cambridge area. We also used social media and word of mouth. Those that registered their interest by emailing were sent a Participant Information Form and a Consent Form to consider before committing. Of the thirty-eight people who expressed interest, twenty-two registered and took part. Five were female and seventeen male. The age range was between twenty and twenty-nine, except one participant who was sixty-five. Due to artifacts, complete data from only ten participants, 9 male and 1 female, all between twenty and twenty-nine years of age, were usable in this study.

Stimuli

The software BCI2000 was used to present the sentences and record the EEG with accurate time-locking to the stimulus presentation. The choice of this software was due to exact time labeling of each stimulus cue on the EEG data, thus allowing for time locking and stimulus word identification during analysis without the concern of possible latency errors in labeling the EEG that may arise due to time delays when passing through long cables.

We presented the participants with sentences on a computer screen. The two test sentences were: 'I believe Islamic law is fair.' and 'I believe Islamic law is not fair.' Both were presented for counterbalancing. If the participant's implicit response to the first sentence was that they agreed, it was expected that the participant's implicit response to the second sentence would be that they disagreed.

These sentences were presented within a longer list of sentences that are not part of the study, but allowed for a distraction to avoid the participant anticipating the theme of the sentences. We also included a third test sentence 'Sharia is not irrational,' Sharia being a wider term for Islamic Law.

Control sentences acted as standards for the study. Four control sentences were presented to the participants:

- Malcolm X was a Prophet
- 7 + 1 = 4
- A car has four wheels
- Mohammed was a Prophet

The first two control sentences were designed to be considered implicitly false by the participants, and the second two control sentences were designed to be considered implicitly true by the participants. The implicit responses to the control sentences were stored.

The implicit responses to a test sentence (e.g., 'I believe Islamic law is fair') could then be compared against the stored implicit responses elicited by the two controls. This allowed us to determine if the implicit response elicited by a test sentence belonged to the category of a 'true' or 'false' implicit control sentence response, as discussed in the Data Analysis section below.

All sentences were presented one word at a time. Each word appeared on a new slide. The slides had a black background, and used white text for contrast. To prepare the slides, JPEG images of each word were made. For each sentence, a blank black slide was inserted between each word to pace the participants equally, and to provide enough time for any impulse under 500ms to have dissipated. Each slide was visible for 500ms. The sentences were structured such that the final word in each sentence was the one that would elicit the implicit response. This response would depend on the reader's implicit attitude. Thus it was only the last word that was expected to generate an N400.

At the end of each sentence, where the N400 was due to be measured, two blank screens were presented to allow for enough time for the signal to be generated. The next slide displayed a question mark. Upon seeing the question mark, a participant, who would be sitting with their arms on the desk, would indicate their answer by tapping

the desk once for 'agree,' twice for 'disagree' and three times for 'unsure' – this would be their 'explicit response.' The act of tapping was found to elicit less interference to the EEG trace than asking them to verbally express their answer. A blank screen then separated the previous sentence from the next. Subjects sat facing a computer screen approximately 70 cm away in a noise-attenuated room to minimise distractions. The sentences were presented in the same order to the participants and were presented once. All participants were exposed to the same sentences.

The study was organized as part of an unpublished graduated MPhil dissertation at the University of Cambridge in a collaboration between the Faculty of Divinity and Department of Clinical Neurosciences.

Data Collection

The EEG of each participant was recorded using a g.tec USBamp (24-bit 16-channel biosignal amplifier, g.tec Medical Engineering GmbH, Austria. Serial Number UA-2007.04.24) at a sampling frequency of 256 Hz. The ground electrode was located on the right mastoid. The reference was selected as the electrode at AFz. Passive gold electrodes were placed on an EEG cap at sixteen recording sites distributed over central and parietal areas where the N400 is known to be maximal (FCz, Fz, F3, F4, Cz, C1, C2, C3, C4, Pz, P3, P4, O1,O2, PO3 and PO4). SuperVisc High Viscosity Electrolyte Gel (EasyCap GmbH, Germany) was used to improve conductivity between the scalp and the electrodes. The impedance values were kept under 25kOhm, and were typically 10k Ohms.

Data Analysis

The original reference was maintained. The EEG recorded was filtered using a digital FIR filter between 0.5Hz and 20Hz. This frequency range captured the N400 ERP. Periods with ocular artifacts in the EEG recordings were removed by visual inspection. The data were segmented into 1000-msec EEG epochs, starting 200-msec before the precise time of onset of the target word (the last word in each sentence) and ending 800-msec after the onset. The epochs were baseline corrected, using the average of the 200-msec pre-stimulus period.

The control sentence implicit response ERPs were averaged. A time window that provided the greatest distinction between the averaged 'false' control implicit responses of all the participants and that of averaged 'true' control implicit responses of all the participants was found. This was between the data points of 409ms – 503ms, on C2 (t-test 2.12, $p=0.003$). This provided for two control groups, an implicitly 'true' control

response, and its opposite, an implicitly 'false' control response. The EEGLAB toolbox for MATLAB (Delorme and Makeig 2004) was used for these analyses.

In order to be able to systematically categorise the implicit response of a participant to a test sentence (e.g., 'I believe Islamic law is fair') as being implicitly true or implicitly false, the implicit test sentence response was compared with implicit responses obtained from control sentences. This comparison was carried out as follows.

To determine which of the two control groups ('true' or 'false') an implicit response elicited by a test sentence belonged to, we employed a statistical comparison. We compared the implicit response elicited by a test sentence against the two control groups using a two-sample t-test. In order to determine if the implicit response elicited by a test sentence was the same or different to either of the two controls, we considered the p value that resulted when the t-test was conducted on the test sentence implicit response data points and each control implicit response data points. The t-test being made separately for each of the two controls against the implicit response elicited by a test sentence. We found that, for all results, the implicit response elicited by a test sentence could be attributed as being the same as either one of the two controls. The outcome of the t-test for each test sentence was a $p < 0.05$ for exclusively one of the two controls. Never were the data points elicited from a test sentence statistically the same as both controls. Each implicit response elicited by a test sentence could thus be categorised as belonging exclusively to either one of the two implicit control groups to a $p < 0.05$.

With all the test sentences now categorized as either implicitly true or implicitly false, we considered whether the implicit responses of the participants were in line with their explicit responses.

Results

Control Sentences

All twenty two participants were presented the four control sentences. To view the data from these, we present the full data set graph plot in Figures 1 & 2.

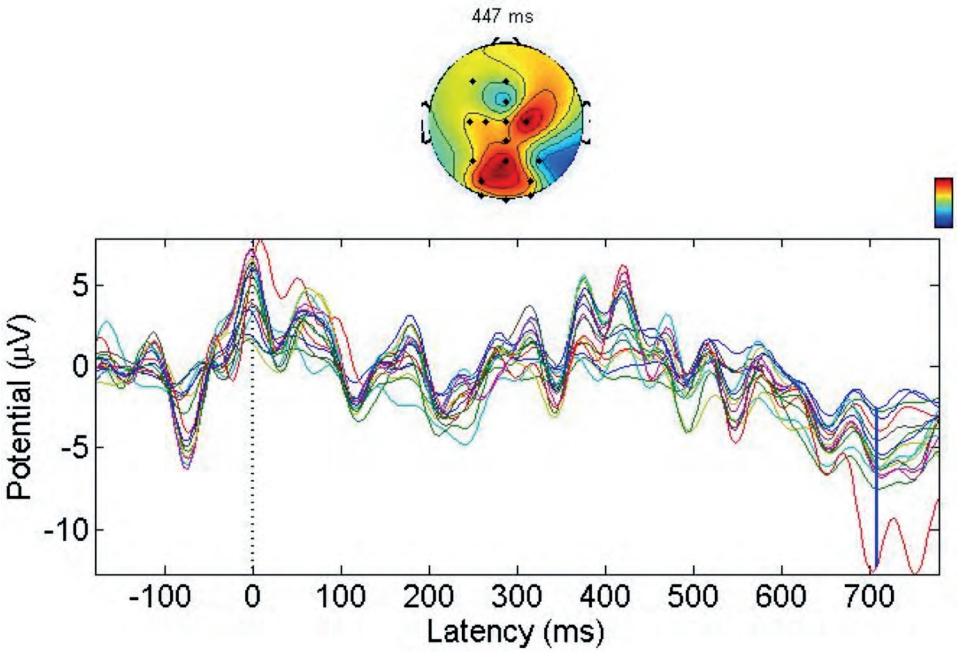


Figure 1. Average of all ERP signals from participants when they read the 'true' controls - sentences that they implicitly agreed with. The N400 wave is minimal (seen between 400ms to 500ms). Epoch from time (t) = -200ms to 800ms – whereby the final word of the control (stimulus onset) occurs at time (t) = 0ms. Each trace represents a channel.

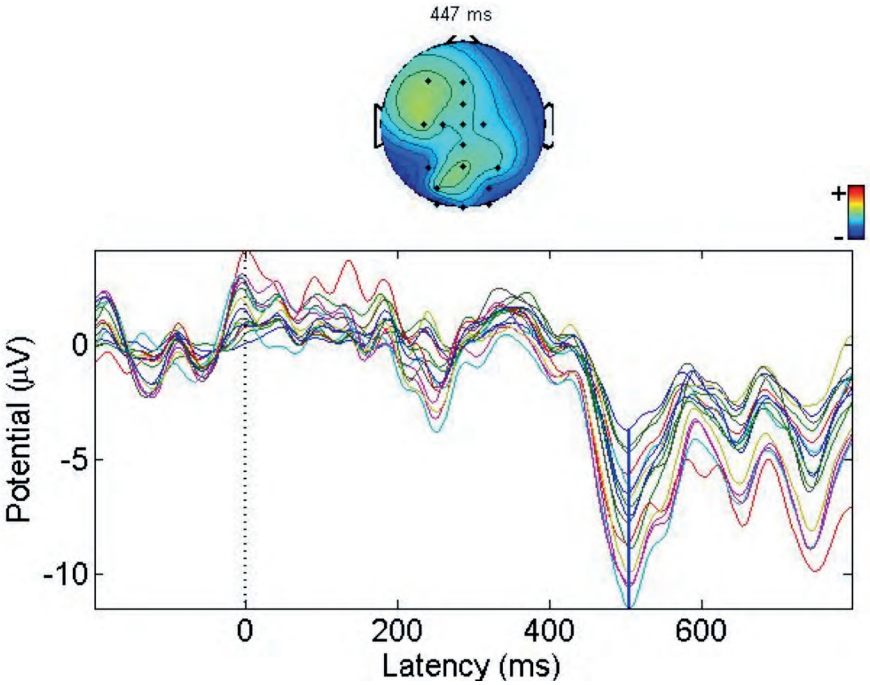


Figure 2. Average of all ERP signals from participants when reading 'false' controls - sentences they were implicitly disagreed with. The N400 wave is maximal (seen as the dip between 400ms to 500ms). Epoch from time (t) = -200ms to 800ms – whereby the final word of the control (stimulus onset) occurs at time (t) = 0.

Figure 1 depicts the responses to 'true' controls, and figure 2 depicts the 'false' control responses.

The results from asking participants whether they agreed or disagreed with the sentence 'I believe Islamic law is fair' is presented next. To visualize these outcomes, we present below the explicit response given by the participants, the implicit categorisation, the averaged implicit score (the ERP averaged over the 409ms – 503ms time window), and whether the implicit result was consistent with their explicit response.

Participant Number	Participant's Explicit Response to: 'I believe Islamic law is fair'	Participant's Implicit Response to: 'I believe Islamic law is fair'	Participant's avg. Implicit score (μV) to: 'I believe Islamic law is fair'	Comparison of Explicit and Implicit result
1	True	False	-2.9778	Inconsistent
2	True	True	0.3508	Consistent
3	True	False	-1.0719	Inconsistent
4	True	False	-5.1088	Inconsistent
5	True	False	-7.4296	Inconsistent
6	True	False	-57.5302	Inconsistent
7	True	False	-4.5338	Inconsistent
8	True	True	1.3363	Consistent
9	True	False	-41.005	Inconsistent
10	True	False	-7.0499	Inconsistent

Table 1. Explicit and implicit responses of all participants to the test sentence 'I believe Islamic law is fair.'

The test of the counterbalancing stimuli sentence 'I believe Islamic law is unfair' is given next.

Participant Number	Participant's Explicit Response to: 'I believe Islamic law is unfair'	Participant's Implicit Response to: 'I believe Islamic law is unfair'	Participant's avg. Implicit score (μV) to: 'I believe Islamic law is unfair'	Comparison of Explicit and Implicit result
1	False	True	9.8057	Inconsistent
2	False	False	0.8198	Consistent
3	False	True	1.1513	Inconsistent
4	False	True	0.7443	Inconsistent
5	False	True	2.0114	Inconsistent
6	False	False	-8.119	Consistent
7	False	False	-0.2034	Consistent
8	False	False	-4.658	Consistent
9	False	False	-3.7474	Consistent
10	False	False	-0.2928	Consistent

Table 2. Explicit and Implicit responses of all participants to the test sentence 'I believe Islamic law is unfair.'

The first result to appear in this determination was that, while all participants explicitly indicated that they agreed with the statement 'Islamic law is fair,' eight participants appeared to disagree implicitly. The two other participants displayed no inconsistency in their results. They explicitly and implicitly believed Islamic law was fair. They were participants 2 and 8. From their questionnaires, they both had attended the same seminary, both believed all areas of religion are meant to be reasonable and both also expressed that there were higher objectives behind Islamic law, both had undertaken an in-depth study of Maqasid (Higher objectives of Islamic law, which is also known as the spirit of the law approach). Both participants 2 and 8 were Maturidi Sunnis. They also both said 'yes' to the question 'Do you believe what is 'good' and 'bad' is naturally known to humankind?' though they disagreed on: 'If humans never had prophets, could they live fairly and prosperously?' Participant 2 answered 'yes.' Both were male and of age 20. Both expressed that 'Right and wrong can depend on the situation.' These two participants contrasted with the other participants in the following way:

These two participants, who were consistent -believing Islamic law was fair both implicitly and explicitly, differed with the remaining inconsistent participants in one educational area: The two consistent participants had undertaken an in-depth academic study of the higher objectives of Islamic law (Maqasid) also known as the spirit of the law approach, as mentioned. An approach that considers fairness and purpose as essential parts of law and law making (Auda 2008, 3, 4, 16, 34, 49).

In considering the responses of all participants to the statement, 'I believe Islamic law is fair' and 'I believe Islamic law is unfair,' we expected that the implicit results of the participants would show that they implicitly agreed with one but not the other. This was the case for participants 1, 2, 3, 4, 5 and 8. These may be considered correctly counterbalanced. This was expected, as an implicit response to one sentence should be the opposite of the other. However four other participants – 6, 7, 9, 10 appeared not to provide counterbalanced results. We consider these next.

For three of those four participants: participant 7, 9 and 10, we note that their implicit responses may be considered counterbalanced if we were considering a comparison of the relative magnitude values for the two opposite sentences. All implicit responses to the first sentence using 'fair' are many orders of magnitude larger than its opposite sentence 'unfair.' To visualize this we present the magnitude ratios below in table 3. The level of the N400 for the test sentence: 'Islamic law is unfair' is many orders smaller than the N400 for the opposite sentence. This indicates an aversion to the 'fair' claim that is many times more substantial than that to the 'unfair' claim. This

may indicate that they implicitly disagreed with the sentence 'I believe Islamic law is fair' to a far higher degree than its opposite. With Participant 6 however, given the high negativity for both sentences, despite being opposite sentences, we may consider that their response may be that of an outlier. Further studies on whether the N400 can be used as a scale of 'agreement' would be beneficial for the larger study to come - as will be outlined at the end of this paper.

Participant Magnitude ratio of the avg. Implicit results to both sentences: (Unfair/ Fair)

Participant 7	0.203/4.534= 0.04%
Participant 9	3.747/41.01 = 0.09%
Participant 10	0.293/7.05= 0.04%
Participant 6	8.119/57.5302 = 14.1% (outlier)

Table 3. Relative average responses to the two sentences 'I believe Islamic law is unfair / fair.'

The second result of this study found that all participants held the implicit belief that Islamic law was unfair despite explicitly stating the opposite, except for two participants, participants 2 and 8. These two participants believed Islamic law was fair at both an implicit and explicit level.

Before discussing the reasons for this inconsistency in the implicit and explicitly expressed outcome, we consider the outcome of presenting the sentence 'Sharia can never be irrational.'

Third Test Sentence Results

Instead of using the term 'Islamic law,' we used the term 'Sharia,' Sharia being a wider term used to describe Islamic law. We replaced 'fair/unfair' by the term 'irrational.' Thus the test sentence became: 'Sharia can never be irrational.' For visualization, the test responses are given in table 3.

Participant Number	Participant's Explicit Response to: 'Sharia can never be irrational'	Participant's Implicit Response to: 'Sharia can never be irrational'	Participant's avg. Implicit score (μV) to: 'Sharia can never be irrational'	Comparison of Explicit and Implicit result
1	False	True	7.2174	Inconsistent
2	False	True	1.526	Inconsistent
3	False	True	2.2694	Inconsistent
4	False	True	8.2501	Inconsistent
5	False	False	-0.0433	Consistent
6	False	False	-7.7214	Consistent
7	Unsure	False	-18.3416	Unknown
8	False	False	-1.6594	Consistent
9	False	False	-38.1368	Consistent
10	True	False	-0.4617	Inconsistent

Table 4. Explicit and implicit responses of all participants to the test sentence 'Sharia can never be irrational.'

All of the participants explicitly expressed disagreement with the sentence 'Sharia can never be irrational' through tapping except for one participant (participant 10), one other remained unsure (participant 7). However, of the participants who expressed explicit disagreement, four of them implicitly agreed with that sentence: Participants 1, 2, 3 and 4. That is, they implicitly held that Sharia can never be irrational (i.e., Sharia is always rational). Three of these participants (1, 3, 4) also implicitly held that Islamic law was unfair (table 1).

Here we pose the question: 'why was it that they implicitly held Sharia can never be irrational, yet also implicitly held that Islamic law can be unfair?' To answer this question we considered the questionnaire responses to see how the participants reasoned on questions such as the trolley problem (deontological, utilitarian, etc.) and it was found that participants 1, 3 and 4 had taken the view that it was rational to push one person off a ship that was sinking due to being used over its capacity in order to save the other passengers. This outcome may indicate that the participants (1, 3 and 4) thus held that Sharia is, form a consequentialist point of view, rational yet unfair.

A main finding thus appears to be that these three participants considered 'what is rational can be distinct to being fair.' It would seem to indicate that in their reasoning, laws that can be rationalised may not necessarily be fair. Such a view, the unfairness yet rationality of Islamic law, would be considered at odds with Islamic legal thought,

however, it appears that it was implicitly held by these participants. This being the case, it may highlight how individuals may hold implicit values that they themselves are unaware of, values that are at odds with their own explicitly expressed value system. This is not an uncommon finding as the Implicit Association Test, a psychological test method that uses timing to compare the strength of association, has shown that many individuals hold biases towards races and other social categories that they themselves were not consciously aware of (Nosek, Banaji, and Greenwald 2002).

We consider next why it was the case that a number of the participants (eight) exhibited an inconsistency between their implicit and explicit results. This inconsistency may be considered a form of cognitive dissonance as we shall explore. Further we will consider why these implicit responses may have materialised to begin with.

Discussion

Why the dissonance between implicit and explicit responses?

Cognitive dissonance is a negative drive state that occurs whenever an individual simultaneously holds two cognitions, be they ideas, beliefs, or opinions, which are psychologically inconsistent, whereby the opposite of one cognition follows from the other (Berkowitz 1978, 2). When an implicit response is found to be different to an explicit response, this may be an indication of cognitive dissonance.

For a Muslim to explicitly suggest that Islamic law is unfair, they could potentially be distancing themselves from their own group. This being too averse a step to take due to group social and emotional attachment, dissonance may result. An alternate reason for the indication of dissonance may be rooted in the finding that there is a tendency to justify the current wider social order even if the status quo goes against one's personal interests (Jost, Banaji, and Nosek 2004). A common finding in the literature on system justification is that members of disadvantaged groups often adopt a negative stereotypical view of their ingroup, thereby protecting their beliefs about the fairness of the current wider social structure. From a cognitive consistency perspective, one may consider that such reactions have their roots in the conflict between the general belief that the existing social structure is fair and the specific belief that one's ingroup is disadvantaged. To the extent that individuals are motivated to retain their general belief about the fairness of the current system, they may restore consistency by adopting the belief that the ingroup is inferior (Gawronski 2012). This may also be a factor that helps to explain why some who identify with religious establishments that ought to

champion fairness, have in the past lent support to authoritarian regimes and unjust social structures instead of lending support to the ingroup that is being treated unfairly.

In addition, it may be that identifying Islamic law as unfair would be the same as suggesting that the 'way of life' of the Muslim community in which they lived was at fault. Protecting their beliefs about the fairness of the current Muslim social structure may have thus also led to the dissonance indicated by the study.

An alternative reason for the indication of dissonance may be as suggested by studies that consider cognitive dissonance to be in part due to 'ego-defense.' Consistency as a core motive for dissonance has been documented in cases related to mechanisms of ego-defense in justice settings (Konow 2000). It may be that maintaining one's view to be correct could have caused such dissonance.

Yet, despite these theories on possible causes of cognitive dissonance, neuropsychological work has demonstrated that dissonance in general might not always be a conscious strategic process (Lieberman et al. 2001). Anterograde amnesia patients, who had neurological damage affecting the functioning of medial temporal lobe and were incapable of forming new memories, were compared with healthy controls on a dissonance task. The amnesics had no memory of having performed a behavior that conflicted with their previously established attitudes and thus were not likely to have engaged in conscious strategic attitude change. Nonetheless, the amnesics changed their attitudes to the same extent as controls. These results suggest that, rather than conscious rationalisation, cognitive dissonance reduction may sometimes depend on implicit constraint satisfaction processes (Read, Vanman, and Miller 1997; Lieberman 2006).

Decisions emanating from implicit-explicit cognitive dissonance

Since the occurrence of dissonance is presumed to be unpleasant, individuals strive to reduce it by adding 'constant' cognitions or by changing one or both cognitions to make them 'fit together' (Berkowitz 1978). One of the ways cognitive dissonance is alleviated, is through rationalisation. Cognitive dissonance theory proposes that the agent is motivated to reduce this tension and may, in this context, do so either by reducing self-interested behavior, or by engaging in self-deception, or by some combination of the two. It is documented as a 'psychological' need (Festinger 1962). It is also seen in children when they seek a form of cognitive consistency (Egan, Santos, and Bloom 2007).

In such circumstances, when individuals perform a behavior or make a choice that conflicts with a previously established attitude, the attitude tends to change in the direction that resolves the conflict with the behavior. This process appears to involve a

rationalisation whereby individuals strategically change their attitudes in order to avoid appearing inconsistent (Jarcho, Berkman, and Lieberman 2011). Of the eight participants who displayed an indication of cognitive dissonance, believing Islamic law to be unfair at an implicit level, three appear to have rationalised their position on this. The study found that the three of the eight had held at an implicit level that 'Sharia was not irrational.' Thus, it appears that they had found a method to accept Islamic law as being a valid and true system of law by rationalising it's perceived unfairness. They did so by ascribing to a form of utilitarianism. To them, it was unfair to push a person off a boat to save the remaining passengers, yet it was the rational course of action. In the same way, it appears, they may have found Islamic law, or elements of it to be unfair, yet justified this by considering such framing as a rational and not irrational position to take on law.

Why did a conflicting implicit attitude manifest?

Irrespective of the method by which participants came to express a view that was contrary to their implicit attitude, there remains a further question. Why did eight of the participants have an implicit attitude that was at odds with Islam, the 'authority' that they believed in?

It may be argued from an empiricist perspective, one that holds 'values' as learned, that a 'value system' of any authority that is subscribed to and practiced would not cause any cognitive dissonance. This would be because the subscriber has submitted themselves to the worldview set by the authority. The values of the person are shaped by the said authority. Such was the case with the participants of this study. This was displayed in their anonymized confidential feedback questionnaire, none of the participants expressed they believed in an alternative worldview. Yet, for eight of the participants, the majority – seven of whom were seminary graduates and involved in religious teaching – it appears there was a factor that perturbed this attitude, one which may have led the mind to a form of cognitive dissonance. Given that – we suggest – it was not an alternate system of beliefs, it may be that the factor that caused this perturbation was actually an innate sense of both fairness and purpose. For without such an innate sense, where else may they have a point of reference that contradicted their system of belief? This hypothesis is also made based on additional information found in this research: The consistency of the implicit and explicit result found with the two participants who subscribed to the Maqasid legal school of law. The school places a determination of fairness and purpose as recourses that are essential to law (Auda 2008, 3–4, 32). We expand on this further, next.

In among the participants were two who had undertaken an in-depth academic study of school of Maqasid, also known as the spirit of law school. The school places authority in the concepts of fairness and purpose, instead of relying on a literal reading of the source texts (Auda 2008, 3–4,32; Ashur 2006). This allows a jurist to change textually mandated laws when a context changes. A law that is fair in one context may not be fair in another, in part because the purpose of the law is no longer met. Consideration of the fairness and purpose of law is central to this school. Law is not seen as dogmatic or irrational, but open to reason. The law's main intentions are seen to revolve around protecting and upholding people's interests, the acquisition of benefit, and mitigation of harm. Early legal scholars such as al-Shatibi articulated these in five main objectives, the protection of life, religion, progeny, wealth, and intellect. These objectives collectively represented the telos to which reasoned deliberation in the law must aim (Emon 2010, 116).

This epistemology of law allows for recourse to fairness and purpose, compared to non-Maqasid schools which rely solely on the text, its authority, in a more literalist approach to law (Jackson 2006). Thus, it may be the case that the Maqasid school allowed a participant's attitude towards law not to clash with their innate human expectation that law needed to be both fair and of purpose. The school may have allowed for a human expectation of fairness and purpose in law to remain unfettered and to implicitly manifest.

Furthermore, the two participants were of the Maturidi theological school that takes a nativist approach to morality, suggesting that good and bad can be recognised without recourse to source texts. However, other participants also adhered to this theological school. It may thus be put forward that, without recourse to a legal mechanism (i.e., Maqasid) that considered fairness and purpose as essential to law and law making, a clash with an innate, nativist expectation occurred. The Maqasid legal school holds a practical methodology. Thus, it appears that in taking a Maturidi stance in theology without a Maqasid practical stance on law, an innate sense of fairness and purpose could not find a practical method to express itself, resulting in cognitive dissonance.

In the case of those participants who did not subscribe to the Maqasid school, and thus effectively subscribed to more literal approaches, placing far more authority in the text, studies have found that such authority lends itself to a form of legal formalism, one where the law appears to the person holding this schema as complete and univocal (Lyons 1998, 258). It has also been found that those holding such attitudes, whereby law is seen as unchanging, exaggerate the role of the text and minimize the role of the human agent who interprets it (Fadl 2009, 98). The more literalist the approach to law

is, the less the concern with its consequences – so far as the wording has been enacted irrespective of its context.

In essence, what the overall findings may thus be suggesting, is that 'there are innate qualifications of fairness and purpose' and that these 'continue to persist at an implicit level despite an individual's subscription to an alternative set of values.'

What may be put to challenge this thesis is that, despite the majority's commitment, practice and seminary learning, the negative implicit results were not due to a clash with an innate sense of fairness and purpose, but were due to a clash with a unconscious learned value system that they unconsciously subscribed to. One that they were not fully aware of. One that considered Islamic law to be unfair and without tangible purpose. However, such would need to be substantiated with the results of the participant's belief on this topic, and the current study found that all the participants believed in the prophethood of Muhammed and truth of his message.

An alternative challenge may be that a person's own ego could be a cause for cognitive dissonance. Hence, it may be that a subject's own egotistic aversion to act fairly alters their implicit response towards such a question as 'I believe Islamic law is fair.' If this could be indeed established, then an additional or alternative hypothesis for the results may be that 'the state of a subject's ego will reflect in their implicit data' – whereby the eight who had negative implicit results had not 'sufficiently controlled their ego' to be at ease with the concept of acting fairly, compared to the two who had 'controlled their ego' sufficiently to be content with acting fairly. While Islam and it's Sufi dimension has within its approach a method to assist an individual to overcome their ego, one that would have an effect on character, the challenge with this theory is that, as it currently stands, the N400 ERP has demonstrated itself as a method by which it establishes core knowledge and belief violations as detailed above, and not the detection of the factor – be it confounding – of a person's own ego.

Innate qualifications of law?

Two elements that make up law are undoubtedly fairness and purpose. Whether human beings are naturally good, bad, or neither, has been a starting point upon which legal philosophers have built their theories, particularly those which relate to social contracts.

In developmental cognition terms on teleological reasoning, an attempt to reduce children's broad teleological bias was carried out in a study that attempted to introduce a pre-trial that described, in non-teleological terms, the physical process by which non-

living natural kinds form. In spite of this attempt, the study replicated the effects of an earlier study in which no pre-trial information was given as to the reasons behind these physical processes (D. Kelemen 1999). It has also been found that young children are prone to generating artifact-like teleo-functional explanations of living and nonliving natural entities and endorsing intelligent design as the source of animals and artifacts (Deborah Kelemen and DiYanni 2005). The same study also revealed that children's teleo-functional and intelligent design intuitions about natural phenomena are interconnected. Indeed the authors opined that children's teleo-functional intuitions might reflect an indefeasible, innate, cognitive bias. This tendency becomes more selective as children acquire increasingly coherent beliefs about causal mechanisms (Lombrozo, Kelemen, and Zaitchik 2007). Teleological reasoning has also been shown with one-year old's (Gergely and Csibra 1997). At around ten to twelve years of age, the preference for teleological explanations lessens (Cruz and Smedt 2015). Yet a preference for teleology persists throughout life, with a distinct developmental continuity observable of a preference for teleological explanations (Lombrozo, Kelemen, and Zaitchik 2007) leading some to put forward the view that teleological reasoning may be innate (Deborah Kelemen and DiYanni 2005).

On fairness, capuchin monkeys have demonstrated a strong aversion to its absence in food share amongst its peers (Brosnan and de Waal 2003) (Lakshminaryanan, Chen, and Santos 2008). In humans, Hamlin has shown an expectation of fairness and an version to unfair behavior in 6 and 16 month old babies (J. K. Hamlin 2015). They contend that although active prosocial behaviors emerge after birth, they are unlikely wholly the result of brute socialization: They find that they occur spontaneously, are present in primates, and are intrinsically motivated (Aknin, Hamlin, and Dunn 2012). By the end of a child's first year, infants categorize goal-helping as positive and goal-hindering as negative. Like adults, infants appear to evaluate others as good and bad mentalistically: 'Good guys are those who knowingly and intentionally facilitate a third party's goal' (Kiley Hamlin et al. 2013).

The expectation of fairness in infants also appears to be projected onto others. Infants have been shown to expect that individuals, treated fairly and unfairly in a resource distribution task, would prefer the fair distributor (Geraci and Surian 2011). Infants from as early 3 and 4.5 months of age have an aversion to hinderers over helpers (Kiley Hamlin, Wynn, and Bloom 2010; J. Kiley Hamlin 2013). Hamilin thus concluded that from extremely early in life, human infants show morally relevant motivations and evaluations — ones that are mentalistic, nuanced, and do not appear to stem from socialization or morally specific experience.

Amazonian Ultimatum Games and the universality of fairness in humans?

It also appears that the characteristic of fairness is universal (Brosnan and de Waal 2003). Moral judgements activate brain regions that are involved in mentalising, including the medial frontal cortex. A study that compared patients who suffered lesions in this region at a young age and those whose cortex was damaged in adulthood found patients with childhood lesions presented with defective social and moral reasoning, whereas this was not evident in those with later damage (Anderson et al. 1999).

To investigate the possible variations on how resources are shared from culture to culture, a study found that a tribe in the Amazon made do with smaller shares in ultimatum games (Henrich 2000), which they suggest may demonstrate that the qualification of fairness is different from culture to culture. This was repeated in fifteen different societies with similar variation in the amount of money offered in the Ultimatum Game (Henrich et al. 2001). Yet, these studies regrettably did not ask the question 'would the individuals taking the larger share and offering the lower share have wished the same for themselves?' It seems the study missed a critical factor in it's consideration of the outcome. Consideration of how others are similar to oneself is of prime importance in hypothesizing the outcome of these games. This has been demonstrated in more recent social cognitive neuroscience research that involve economic exchange with social dynamics. These studies use paradigms such as the Ultimatum Game (Sanfey et al. 2003) to examine the neural responses associated with cooperation, competition, fairness, and trust. Across these studies, cooperation, trust, and fair play typically activate the VMPFC, MPFC, and MPAC (Decety et al. 2004; McCabe et al. 2001), whereas unfair and untrustworthy responses activate insula (Sanfey et al. 2003), caudate in the basal ganglia (de Quervain et al. 2004), or DMPFC (Decety et al. 2004). The finding that cooperation, relative to competition, promotes MPFC rather than DMPFC activity, is also consistent with previously described work (Mitchell, Banaji, and Macrae 2005) such that cooperation may be associated with seeing the other players as more similar to oneself, since cooperation, relative to competition, promotes MPFC rather than DMPFC activity (Lieberman 2006).

This is consistent with our stipulation that a study on fairness in ultimatum games ought to pose the question 'does the player wish the same share for themselves.' Considered symmetrically: If the opposite player were themselves, would they offer the same share? Such a question would allow us to assess whether or not the player views the other person as similar to themselves, in which case, the above social cognitive neuroscience research appears to suggest that the player would more likely be fair to them.

Across these studies on fairness and trust, the fairness of the decision-making process has often been confounded with the material value of the outcome. That is, fair responses from a partner are typically associated with better financial outcomes for the subject. Tabibnia recently manipulated the material payoffs and the fairness of the partner's behavior independently. After controlling for material payoffs, fairness still activated an array of motivation- and reward-related regions, including the VMPFC, ventral striatum in the basal ganglia, and amygdala, which suggests that fairness is hedonically valued in social interactions (Lieberman 2006).

We also suggest that asymmetrical relations can belie a self-centered cognition. Thus, exploitation ought to be considered as a behavioral factor to be measured in ultimatum type studies, particularly when a society is scarce of resource and individuals can often 'make do' with that they receive. The knowledge of this circumstance can spur a person in power to take advantage of the situation and offer less than they would outside of this context.

Context: Law, morality and cognition

Within early Muslim jurisprudence, two approaches to law existed. One was based on source texts, whereby the source texts defined justice, compassion, benefit and harm. The enterprise took to an empiricist type of approach towards morality. The second school of law saw in the source texts representations of justice, compassion, the acquisition of benefit and mitigation of harm. This form of reasoning also manifested itself in a specific method for arriving at law, a method that became known as istihsan (juristic preference). It was driven by reasonableness, fairness, common sense and public interest set as deriving the most good and mitigating the most harm, both of which involved reasoning that did not appear to be directly based on the source texts (Izzidien In Press; Hallaq 2005, 116). Maqasid began to be developed as a bridge between the two approaches to law (Izzidien In Press). The theological stance of the Maturidis, Atharis and Mutazilites allowed for a rationalisation of law that set it's purpose as being the benefit of humankind, set in fairness – not one based in dogma and literalism. In the current twenty-first century, Muslim Democrats have taken to Maqasid, seeing it as an authentic means for the re-interpretation of law to allow it to remain dynamic and suited to new contexts (Glancy 2007, 35). Some recent reformers have, it has been suggested, attempted to use Maqasid as a means beyond what it was intended for (Emon 2010, 188). How law is articulated today in many countries in the Muslim World have their

roots in these early discussions found in Islamic legal theory. These discussions continue today in similar lines to those on legal formalism and realism.

Follow up studies – two new hypotheses.

With the inductive part of the study complete, we are now able to set two new hypotheses to be examined.

The first hypothesis would be 'Ascribing to the spirit of law school of thought (Maqasid) exclusively produces an implicit belief that Islamic law is fair.' To test this hypothesis we would need to recruit two larger groups of participants, one that ascribes to this school and a second that specifically does not. Larger groups will allow the study to be more representative. It may also be useful to consider, in a third study, whether the implicit views of the latter school could be changed – if any of them were to want to reconsider their views after the study.

The second hypothesis would be on the 'nativism and empiricism' area of study in this paper. While most studies on this revolve around developmental cognition, given that babies and infants have yet to be 'socialised' into taking up a set of values, providing ideal for research into such studies, we propose an alternative method usable in adults.

Given the empiricist approach holds that values held are always learned, we would seek to find adults who subscribe to a named value system (or authority), implicitly and explicitly, be it religious or non-religious. Of these adults we would attempt to detect any discrepancy between their expressed views and implicit views on the values associated with that named value system (or authority) (e.g., it's fairness). Where such a dissonance occurs we would seek to decide if such a dissonance could be related to an innate factor or to an alternate socialised learned factor using implicit values measures. The study would ask if it is indeed possible for an adult to hold implicit values that contradict nativist theories on innate values. These methods could naturally be extended to other forms of concepts. The measurement of implicit responses to a wide array of values would be necessary in order to remove confounding factors.

This study may be further extended to the study of law in general and that of judgments made by judges who subscribe to an ideology, as their judgments may also be unconsciously biased. This has been found in research on bias, political ideological subscription and court case outcomes in non-implicit data research on the topic (Sunstein et al. 2007).

Further, fMRI data on moral decisions has shown that moral problems given their personal dimension, activated a medial frontoparietal network along with LPAC to a

greater degree than an impersonal condition, consistent with notion that a personal condition promotes self reflection on the implications of one's contribution to the outcome. Whereas an impersonal condition, in contrast, leads to greater activity in lateral frontoparietal regions than does the personal condition, consistent with an external focus on events in the world (Greene et al. 2001) (Mendez, Anderson, and Shapira 2005) (Lieberman 2006). It may thus be useful to investigate how individuals with the indications of dissonance found in this study, those who rationalised their moral decision to push a person of a ship to save the others, against those who did not rationalise their decisions, compare under fMRI during these evaluations.

Conclusion

In this study, we sought to detect the implicit beliefs of participants towards the fairness of the law that they followed, in this case Islamic law. Eight participants who had not studied the Maqasid spirit of law school and who were only familiar with a school of law that places an emphasis on the religious texts above that of a determination of fairness and purpose, appeared to exhibit a conflict between what they implicitly believed and what they claimed to believe. It may be that this conflict arose because their epistemology of law did not coincide with a theorised innate sense of fairness and teleological reasoning. If aspects of their law were perceived as unfair and lacking purpose, yet their legal school opposed considerations of these perceptions, then a form of dissonance, between explicit and implicit beliefs as found in the paper, may have been the outcome. Further, a person with such may have had to seek alternate ways to ameliorate this inconsistency. This may have taken the trajectory observed with those participants in this study who differentiated between that which is rational and that which is fair. Indeed a more literalist approach to law can be less concerned with the outcome than loyalty to the meaning of the texts used in law. The study of the cognition of law may offer us new perspectives on legal perceptions, especially those that seek to make fair judgment, be it through social contracts or non formalised legal systems.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

AI proposed the study and its content; AI & SC designed the study; AI conducted the experiment; AI analyzed the data; AI wrote the manuscript, SC contributed to the editing of the manuscript.

References

Aknin, Lara B., J. Kiley Hamlin, and Elizabeth W. Dunn. 2012. "Giving Leads to Happiness in Young Children." *PLOS ONE* 7 (6): e39211.

Anderson, S. W., A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio. 1999. "Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex." *Nature Neuroscience* 2 (11): 1032–37.

Ashur, Muhammad Al-Tahir Ibn. 2006. *Ibn Ashur: Treatise on Maqasid Al-Shariah*. IIIT.

Auda, Jasser. 2008. *Maqasid Al-Shariah A Beginner's Guide*. International Institute of Islamic Thought (IIIT).

Bargh, John A., and Tanya L. Chartrand. 1999. "The Unbearable Automaticity of Being." *American Psychologist* 54 (7): 462–79.

Berkowitz, Leonard. 1978. *Advances in Experimental Social Psychology*. Academic Press.

Brosnan, Sarah F., and Frans B. M. de Waal. 2003. "Monkeys Reject Unequal Pay." *Nature* 425 (6955): 297–99.

Cropanzano, Russell, Zinta S. Byrne, D. Ramona Bobocel, and Deborah E. Rupp. 2001. "Moral Virtues, Fairness Heuristics, Social Entities, and Other Denizens of Organizational Justice." *Journal of Vocational Behavior* 58 (2): 164–209.

Cruz, Helen De, and Johan De Smedt. 2015. *A Natural History of Natural Theology: The Cognitive Science of Theology and Philosophy of Religion*. 1 edition. Cambridge, Massachusetts: MIT Press.

Decety, Jean, Philip L. Jackson, Jessica A. Sommerville, Thierry Chaminade, and Andrew N. Meltzoff. 2004. "The Neural Bases of Cooperation and Competition: An FMRI Investigation." *NeuroImage* 23 (2): 744–51.

Delorme, A, and S Makeig. 2004. "EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics." *Journal of Neuroscience Methods* 134: 9–21.

Egan, Louisa C., Laurie R. Santos, and Paul Bloom. 2007. "The Origins of Cognitive Dissonance: Evidence From Children and Monkeys." *Psychological Science* 18 (11): 978–83.

Emon, Anver M. 2010. *Islamic Natural Law Theories*. OUP Oxford.

Fadl, Khaled M. Abou El. 2009. *The Great Theft: Wrestling Islam from the Extremists*. Harper Collins.

Festinger, Leon. 1962. *A Theory of Cognitive Dissonance*. Stanford University Press.

Gawronski, Bertram. 2012. "Back to the Future of Dissonance Theory: Cognitive Consistency as a Core Motive." *Social Cognition* 30 (6): 652–68.

Gawronski, Bertram, Etienne P. LeBel, and Kurt R. Peters. 2007. "What Do Implicit Measures Tell Us?: Scrutinizing the Validity of Three Common Assumptions." *Perspectives on Psychological Science* 2 (2): 181–93.

Geraci, Alessandra, and Luca Surian. 2011. "The Developmental Roots of Fairness: Infants' Reactions to Equal and Unequal Distributions of Resources." *Developmental Science* 14 (5): 1012–20.

Gergely, György, and Gergely Csibra. 1997. "Teleological Reasoning in Infancy: The Infant's Naive Theory of Rational Action: A Reply to Premack and Premack." *Cognition* 63 (2): 227–33.

Glancy, Brian. 2007. *Liberalism Without Secularism?: Rachid Ghannouchi and the Theory and Politics of Islamic Democracy*. Dublin.

Graham, Jesse, Zoe Englander, James P. Morris, Carlee Beth Hawkins, Jonathan Haidt, and Brian A. Nosek. 2012. "Warning Bell: Liberals Implicitly Respond to Group Morality Before Rejecting It Explicitly." SSRN Scholarly Paper ID 2071499. Rochester, NY: Social Science Research Network.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. "An FMRI Investigation of Emotional Engagement in Moral Judgment." *Science (New York, N.Y.)* 293 (5537): 2105–8.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes." *Psychological Review* 102 (1): 4–27.

Hallaq, Wael B. 2005. *The Origins and Evolution of Islamic Law*. Cambridge University Press.

Hamlin, J. K. 2015. "The Case for Social Evaluation in Preverbal Infants: Gazing toward One's Goal Drives Infants' Preferences for Helpers over Hinderers in the Hill Paradigm." *Frontiers in Psychology* 5.

Hamlin, J. Kiley. 2013. "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core." *Current Directions in Psychological Science* 22 (3): 186–93.

Henrich, Joseph. 2000. "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining among the Machiguenga of the Peruvian Amazon." *The American Economic Review* 90 (4): 973–79.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *The American Economic Review* 91 (2): 73–78.

Izzidien, Ahmed. In Press. "Natural Law and Fiqh." In *Routledge Handbook of Islamic Law*, edited by K. Abou El Fadl and M. Shams.

Jackson, Sherman A. 2006. "Literalism, Empiricism, and Induction: Apprehending and Concretizing Islamic Law's Maqasid Al-Shari'ah in the Modern World." *Michigan State Law Review* 2006: 1469.

Jarcho, Johanna M., Elliot T. Berkman, and Matthew D. Lieberman. 2011. "The Neural Basis of Rationalization: Cognitive Dissonance Reduction during Decision-Making." *Social Cognitive and Affective Neuroscience* 6 (4): 460–67.

Johnson, Russell E., and Robert G. Lord. 2010. "Implicit Effects of Justice on Self-Identity." *The Journal of Applied Psychology* 95 (4): 681–95.

Johnson, Russell E., and Kristin N. Saboe. 2011. "Measuring Implicit Traits in Organizational Research: Development of an Indirect Measure of Employee Implicit Self-Concept." *Organizational Research Methods* 14 (3): 530–47.

Jost, John T., Mahzarin R. Banaji, and Brian A. Nosek. 2004. "A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo." *Political Psychology* 25 (6): 881–919. doi:10.1111/j.1467-9221.2004.00402.x.

Kanfer, Ruth. 2009. "Work Motivation: Identifying Use-Inspired Research Directions." *Industrial and Organizational Psychology* 2 (1): 77–93.

Kelemen, D. 1999. "Why Are Rocks Pointy? Children's Preference for Teleological Explanations of the Natural World." *Developmental Psychology* 35 (6): 1440–52.

Kelemen, Deborah, and Cara DiYanni. 2005. "Intuitions About Origins: Purpose and Intelligent Design in Children's Reasoning About Nature." *Journal of Cognition and Development* 6 (1): 3–31.

Kelemen, Deborah, Joshua Rottman, and Rebecca Seston. 2013. "Professional Physical Scientists Display Tenacious Teleological Tendencies: Purpose-Based Reasoning as a Cognitive Default." *Journal of Experimental Psychology. General* 142 (4): 1074–83.

Kiley Hamlin, J., Tomer Ullman, Josh Tenenbaum, Noah Goodman, and Chris Baker. 2013. "The Mentalistic Basis of Core Social Cognition: Experiments in Preverbal Infants and a Computational Model." *Developmental Science* 16 (2): 209–26.

Kiley Hamlin, J., Karen Wynn, and Paul Bloom. 2010. "Three-Month-Olds Show a Negativity Bias in Their Social Evaluations." *Developmental Science* 13 (6): 923–29.

Konow, James. 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *The American Economic Review* 90 (4): 1072–91.

Lakshminaryanan, Venkat, M. Keith Chen, and Laurie R. Santos. 2008. "Endowment Effect in Capuchin Monkeys." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1511): 3837–44.

Lau, Ellen F., Colin Phillips, and David Poeppel. 2008. "A Cortical Network for Semantics: (De)Constructing the N400." *Nature Reviews Neuroscience* 9 (12): 920–33.

Leuthold, Hartmut, Angelika Kunkel, Ian G. Mackenzie, and Ruth Filik. 2015. "Online Processing of Moral Transgressions: ERP Evidence for Spontaneous Evaluation." *Social Cognitive and Affective Neuroscience* 10 (8): 1021–29.

Lieberman, Matthew D. 2006. "Social Cognitive Neuroscience: A Review of Core Processes." Review-article.

Lieberman, Matthew D., Kevin N. Ochsner, Daniel T. Gilbert, and Daniel L. Schacter. 2001. "Do Amnesics Exhibit Cognitive Dissonance Reduction? The Role of Explicit Memory and Attention in Attitude Change." *Psychological Science* 12 (2): 135–40. doi:10.1111/1467-9280.00323.

Lind, A. 2002. "Fairness Heuristic Theory: Justice Judgments as Pivotalcognitions in Organizational Relations." In *Advances in Organizational Justice*, edited by Jerald Greenberg. Stanford University Press.

Lombrozo, Tania, Deborah Kelemen, and Deborah Zaitchik. 2007. "Inferring Design: Evidence of a Preference for Teleological Explanations in Patients With Alzheimer's Disease." *Psychological Science* 18 (11): 999–1006.

Lyons, David. 1998. "Legal Formalism and Instrumentalism - A Pathological Study." In *Evolution and Revolution in Theories of Legal Reasoning: Nineteenth Century Through the Present*, edited by Scott Brewer. Taylor & Francis.

McCabe, K., D. Houser, L. Ryan, V. Smith, and T. Trouard. 2001. "A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange." *Proceedings of the National Academy of Sciences of the United States of America* 98 (20): 11832–35.

Mendez, Mario F., Eric Anderson, and Jill S. Shapira. 2005. "An Investigation of Moral Judgement in Frontotemporal Dementia." *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology* 18 (4): 193–97.

Metcalfe, Janet, and Walter Mischel. 1999. "A Hot/Cool-System Analysis of Delay of Gratification: Dynamics of Willpower." *Psychological Review* 106 (1): 3–19.

Mitchell, Jason P., Mahzarin R. Banaji, and C. Neil Macrae. 2005. "The Link between Social Cognition and Self-Referential Thought in the Medial Prefrontal Cortex." *Journal of Cognitive Neuroscience* 17 (8): 1306–15.

Nosek, Brian A., Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. "Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site." *Group Dynamics: Theory, Research, and Practice* 6 (1): 101–15.

Poling, Devereaux A., and E. Margaret Evans. 2002. "Why Do Birds of a Feather Flock Together? Developmental Change in the Use of Multiple Explanations: Intention, Teleology and Essentialism." *British Journal of Developmental Psychology* 20 (1): 89–112.

Quervain, Dominique J.-F. de, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. 2004. "The Neural Basis of Altruistic Punishment." *Science (New York, N.Y.)* 305 (5688): 1254–58.

Read, S. J., E. J. Vanman, and L. C. Miller. 1997. "Connectionism, Parallel Constraint Satisfaction Processes, and Gestalt Principles: (Re) Introducing Cognitive Dynamics to Social Psychology." *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* 1 (1): 26–53.

Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science* 300 (5626): 1755–58.

Sunstein, Cass R., David Schkade, Lisa M. Ellman, and Andres Sawicki. 2007. *Are Judges Political?: An Empirical Analysis of the Federal Judiciary*. Brookings Institution Press.

Tripp, Thomas M, Robert J Bies, and Karl Aquino. 2002. "Poetic Justice or Petty Jealousy? The Aesthetics of Revenge." *Organizational Behavior and Human Decision Processes* 89 (1): 966–84.

Van Berkum, Jos J.A., Bregje Holleman, Mante Nieuwland, Marte Otten, and Jaap Murre. 2009. "Right or Wrong?: The Brain's Fast Response to Morally Objectionable Statements." *Psychological Science* 20 (9): 1092–99.

Wiswede, Daniel, Nicolas Koranyi, Florian Müller, Oliver Langner, and Klaus Rothermund. 2013. "Validating the Truth of Propositions: Behavioral and ERP Indicators of Truth Evaluation Processes." *Social Cognitive and Affective Neuroscience* 8 (6): 647–53.

Wittenbrink, Bernd, and Norbert Schwarz. 2007. *Implicit Measures of Attitudes*. Guilford Press.

Journal of Cognition and Neuroethics

Should We Distrust Our Moral Intuitions? A Critical Comparison Of Two Accounts

Felix Langley
Kings College London

Biography
Felix Langley is a recent graduate in philosophy from Kings College London. He specialises in the philosophy of cognitive science and neuroethics with particular focus on the role of intuitions in moral judgment.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). January, 2018. Volume 5, Issue 2.

Citation
Langley, Felix. 2018. "Should We Distrust Our Moral Intuitions? A Critical Comparison Of Two Accounts." *Journal of Cognition and Neuroethics* 5 (2): 51–74.

Should We Distrust Our Moral Intuitions? A Critical Comparison Of Two Accounts

Felix Langley

Abstract

In this paper, I explore the question of whether we ought to defer to our moral intuitions across a range of situations by critically comparing two of the major views on this debate. The views I compare are those of Gerd Gigerenzer and Joshua Greene. Despite both having influential and opposing views, they have never engaged with each other in print and are not often directly compared. Gigerenzer is of the view that our moral intuitions are, broadly speaking, adaptive, whilst Greene takes the opposite view. The main contention that I focus on is Greene's supposition that our moral intuitions are maladaptive in what he calls 'unfamiliar' moral situations (i.e., problems that have arisen in our recent history; e.g., global poverty, terrorism, trolley problems, etc.). My conclusion is that Gigerenzer's thesis is either trivial or false because the areas that Greene identifies as being unsuited to our intuitions are precisely the areas that we should care about, and the conceptual tool that Gigerenzer employs to avoid this (ecological rationality) cannot plausibly solve these problems. The normative framework I employ to judge an intuition as 'better' or 'worse' is one to which both parties can agree.

Keywords

Joshua Greene, Gerd Gigerenzer, Moral Psychology, Intuition, Cognitive Science, Neuroethics, Heuristics, Behavioural Economics

1. Introduction

When one reads any great work of ethics in the western cannon – Kant, Hume, Hobbes, and Aristotle – an account of how we ought to be is intimately connected with an account of how we are. The central thesis of this paper is that across a significant range of moral cases we ought to distrust our moral intuitions. I will argue for this by critically comparing two major contemporary accounts of moral psychology, one which seeks to defend our intuitions and one which challenges them. The accounts in question are that of Gerd Gigerenzer and Joshua Greene; specifically, I will compare the opposing prescriptive claims these accounts make regarding how our intuitions should be employed in moral problems. My conclusion will be that Joshua Greene is correct in his claim that, across a significant class of cases in the moral domain, we should distrust the conclusions of our intuitive moral judgment, and, instead, default to our more reflective moral judgments.

The layout of this paper will be as follows: in the first half, I will explicitly state my normative framework and expound the two major theories, their normative prescriptions, and what the central conflict between them rests upon. In the second half, I will show that Gigerenzer's account fails in it's prescriptive project because: (1) his claims can ultimately be shown to be philosophically trivial; and, (2) the concept that this theory could use to resist this charge, Ecological Rationality, cannot be applied in the relevant cases. By (1), I mean that across what we might think of as 'significant cases (i.e., morally important, high stakes cases) Gigerenzer's account seems to fail, even if it succeeds across more mundane or non-moral cases. By (2), I mean that the concept of ecological rationality, upon which Gigerenzer leans heavily, is an implausible strategy in the cases with which we are concerned.

1.1. Normative Framework

It is important to establish, before moving forward, what exactly will be meant in the following by 'good' and 'bad and 'adaptive' and 'maladaptive'. In order to compare the two theories, we must have some consistent common conception of 'good'. In order to avoid certain moral disputes, both metaethical and normative, the framework I will employ here will be as minimal as it can be whilst still being useful. I will suggest that a feature of our moral psychology is maladaptive if it leads us to make choices based on factors which we would not consciously agree are relevant. What I mean by this is that, for any factor X, if X is shown to effect our choices but we would not when asked agree that X is morally relevant, X can be called maladaptive. Things that would fall into this category would be factors like the order in which information is presented, spatial proximity, and whether a harm was caused physically or remotely – factors, which, when asked, people would say shouldn't be relevant to our decision making. To use a further hypothetical, though as we will see not entirely absurd, example, if it transpired that a person's hight factored into whether we felt it was morally wrong to harm them, then it would seem that this tendency could be called bad or maladaptive since hight just isn't the sort of thing which we, both lay people and ethicists, take to be morally relevant.

In addition to this idea of normative irrelevance, features of common sense ethics like 'all other things being equal inconsistency is bad' and 'all other things being equal it is better that less people be harmed than more' will factor into this framework. These are the kind of normative features which Greene describes as being "uninteresting" (Greene 2014, 771), insofar as they are claims we all seem to accept and do not tend to be the major salient features of moral argument. I accept that many elements of this normative

framework can be subject to metaethical scrutiny. For example, it has been argued that the legitimacy of prevailing our stated preferences over our relieved one's can be called into question. However, I take this framework to be plausible enough that I will assume it to be true for the purposes of this essay, since: (a) even if elements of it can be questioned it is the morality most of us seem to accept and live by, thus it remains an interesting question to see how these competing models fair under it; and, (b) a deeper metaethical argument would take us beyond the scope of this paper.

2. Kahneman, Tversky And Two System Models

The model of decision making being proposed by Greene has its origins in the work of Danial Kahneman and Amos Tversky (1974, 1124–1131). Furthermore, much of Gigerenzer's work is an explicit reaction their project. Thus, in order to correctly understand both Greene and Gigerenzer's accounts, it becomes important to understand the broader approach that informs them. Kahneman and Tversky's project, often referred to as the 'heuristics and biases account', posits a two system model of decision making wherein our minds can 'switch' between a slow, deliberative system of thinking (system two in Kahneman's terminology), and a faster, seemingly effortless, and less deliberative system that makes greater use of heuristics and mental shortcuts (Kahneman 2003, 697–720).

System Two is what seems to allow human beings to solve complex, novel problems and adapt quickly to change (Kahneman 2003, 669). Using it, we can consciously and deliberately focus our attention to a given problem (e.g., solving an equation, choosing a gift for our significant other, coming up with a philosophical argument, and learning how to play an unfamiliar instrument etc.). In short, system two seems to largely capture what we mean when we talk about reasoning or deliberate problem solving. In terms of describing its phenomenal character, we might think of the difference between driving on familiar roads, absentmindedly performing turns in a way which seems automatic but then being confronted with an unfamiliar diversion. The diversion creates a 'cognitive load' which causes us enter into a state of conscious problem solving, calling to mind alternative routes and performing rough calculations of time, perhaps assessing what could have caused it and assigning these causes varying degrees of probability. Similarly, when I arrive at class, the process of walking to the room, sitting down, unpacking my bag, happens almost entirely without any 'input' from me in a way in which I'm aware. However, as soon as the lecturer asks if anyone can see the flaw in an argument that's been presented or knows what school of philosophy Simone de Beauvoir is most

associated with, then cognitive load is applied and my thought requires input in a way that I am aware of and which appear to me more effortful than walking to class etc.

In short, system two is the 'precision tool' of our cognitive architecture. To be clear, I don't mean here that system two is more accurate *per se* or that it isn't capable or even prone to error, rather than is it 'precise,' in that it exists to solve novel, specific problems that require cognitive flexibility and deliberate reasoning. However, obviously, we cannot reason in this way all the time; the trade off we make with system two is that it is difficult and time consuming, hence our reserving it for specific situations where cognitive load is applied. For all other situations, we have System One.

System One is the inverse of the above, trading off precision and adaptiveness for speed and ease. System one thoughts are what we tend to think of as intuitive, reactions and process which come to mind fully formed, and indeed often don't 'come to mind at all'. This is the system which makes greater use of Heuristics. Heuristics, under this account, are (usually unconscious) mental 'rules of thumb' that our brains use in lieu of more complex decision procedures, as they tend to isolate one feature of the situation and make the choice based upon that. Prominent examples studied by Kahneman and Tversky include the availability heuristic, which bases the probability of an event occurring exclusively on how easily one can call to mind an example of it, the representative heuristic, which assesses whether A is a member of class B based on A's approximate resemblance to a mental model or stereotype of B. For example, people, when asked, overwhelming thought that a character with a meek, orderly description is assessed to be more likely to be a librarian than a list of other professions without consideration of other factors, such the statistical distribution of these professions.

There is an important point to make here that is easy to overlook when considering the interplay between these two systems. System Two makes all judgments, but it does not necessarily modify all judgments. This is the difference between judgments made under cognitive load, wherein there is deliberation, and those made intuitively. Thus, if someone where to ask me, on the spot, which has a greater population, Detroit or Grand Rapids, I might intuitively choose Detroit because of the recognition heuristic. If I am asked to think harder about it, system two might engage in deliberation, or if I am not, then I might simply settle on Detroit. This is the difference between an intuitive and non-intuitive judgment, both are made ultimately by system two, but the intuitive judgment is made entirely on the basis of the information that system one provides.

Having laid out the board strokes of the heuristics and biases account, I will now go on to expand why it is termed the 'biases' account'. As is perhaps already becoming clear, for all the duel system's elegance, our propensity to be hugely more sensitive to certain

sorts of information when making choices than to others is likely to have epistemically unhygienic consequences. According to Kahneman's account, as their work in this field progressed they began to increasingly find that these heuristics and mental shortcuts were, across a number of domains, leading us to error on a systemic level. These errors are termed 'biases'. To clarify this picture, I will expound some of the specific features of our intuitions which can cause them to lead us astray:

2.1.Accessibility

The 'pull' that our intuitions seem to have on us, in terms of their appeal as options, appears to stem from the ease with which we can access them. As has been discussed, conscious deliberation to answer a question is slow but our intuitions seems to pop into mind fully formed without (from our perspective) us needing to exert cognitive effort. This appeal of accessibility means that we are likely to, for example, take the frequency of events (e.g., terror attacks or air travel disasters) to correlate strongly with how easily we can bring to mind an example of one occurring. Now, naturally, in a great many everyday cases this heuristic will, in fact, lead us to truth, but it is very easy to see that, across a range of examples, this won't hold true, and, indeed, we might further worry that many of the examples in which it won't hold true are situations that invite dangerous outcomes (e.g., a voting population that demands harsh checks on personal freedom because of a perception that terror attacks are occurring several orders of magnitude more commonly than they are).

2.2.Sensitivity To Framing Effects

A further significant concern is the degree to which our intuitive judgments are sensitive to framing effects (i.e. presenting a problem in a particular way or using some words rather than others leads to substantial changes in the outcomes of choices in a way which we might think is deeply problematic) (Tversky and Kahneman 1981, 453–458). One major example of this effect is that people are highly sensitive to how losses and gains are presented such that the majority of people, when presented with two strategies for preventing a disease, will overwhelmingly choose the less risky option when it is presented in terms of the lives it saves (e.g., will save 200 lives) but overwhelmingly not when it is presented in terms of lives lost (e.g., 400 lives will be lost). This is despite the fact that, in both cases, participants knew the total number of lives at sake. Thus, the only thing that altered most people's choice was a slight change in how the information was framed.

The problem runs even deeper than this though. It is tempting to think that more 'rational,' intelligent people would rely less on heuristics or intuitions and thus be less susceptible to biases; however, this is not the case. Not only does greater intelligence fail to track freedom from bias, but also highly intelligent people are not free from biases in domain specific situations in their own fields. One of Kahneman's studies found that graduate statisticians at one of the top colleges in the US were led to a conclusion by the representative heuristics that basic statistics tells us is impossible (Kahneman 2003, 712).

3. Greene: Automatic And Manual Morality

Now that we have the broader theoretical framework in place, we can discuss Greene's approach from a place of understanding. To a large degree, Greene's account consists in importing the heuristics and biases account unchanged into the moral domain. He explicitly accepts Kahneman and Tversky's model (Greene 2014, 695–726) and shows how it can be cashed out in moral situations. The analogy Greene uses to build on their system one and two notion is the ' Camera analogy'. The camera analogy compares system one to the automatic settings on a digital camera, preconfigured settings which allow the photographer not to worry about adjusting the variables themselves, thus making it highly efficient but inflexible. By contrast, the manual setting allows each individual variable to be adjusted manually, meaning that it is highly flexible but deeply inefficient. Having both these 'settings' make both a camera and a human mind highly efficient.

The central test cases that Greene employs to demonstrate how this model works in practice are 'trolley problems'. Trolley problems (Foot 1967, 5–15) are a broad class of philosophical problems and have been studied since the 1960s as 'test cases' for moral theories and various adaptations and iterations of these moral theories. Like fruit flies in biology, trolley problems are thought to tease out some of the most basic elements of our moral views. The basic problem is as follows: a runaway rail cart (or trolley) is heading down a track towards a group of five workmen and you can save them by pulling a switch, which will change the course of the trolley onto a track that will only kill one workman.

The switch case seems like a relativity easy choice, a trade of one life for five, but now lets say we're faced with a different version of the problem, namely the 'footbridge' case. In the footbridge case, we see a trolley heading towards five workman, but, this time, we cannot pull a switch to stop it. What we can do, however, is push a man off an overhead footbridge into its path, which will save the five workman but kill the one man

on the footbridge. As with the above, this is a one for five trade; however, this strikes us as a much more morally difficult choice, as our intuitions seem to resist pushing the man in a way that they don't resist pulling the switch.

The explanation for this disparity can be easily given in terms of the two system model. The footbridge case, unlike the switch case, involves factors to which our intuitions (and the heuristics that inform them) are sensitive, the most notable in the footbridge case being personal force. As Greene et al describe it (2009, 364–371), it seems that our moral intuitions (i.e., our system one moral judgments) are highly sensitive to the application of physical personal force in a way which 'counts against' the decision being taken.

Multiple studies by Greene et al involving brain imaging; lesion studies, and a variety of self-report studies, seem to bear out this above hypothesis. Cases which are constructed in ways that trigger the sensitivities of our system one tend to produce moral judgments which are: (a) more emotional; and, (b) more associated with what we would typically call 'deontological' moral judgments (Greene et al 2004, 389–400). To unpack this a little, cases in which we are required to push someone off a bridge, or more viscerally smother a child to prevent soldiers being alerted, involve features like direct personal force which trigger an emotional system one response. This then comes into conflict with our more deliberative system two response, often in a way which overrides the cost benefit analysis that characterises this sort of response.

The major contention that Greene draws from this descriptive account, in line with Kahneman and Tversky, is that, in spite of its many adaptive features, system one is often sensitive to features of situations that couldn't possibly be relevant to the decision being taken. We can understand this as a form of 'moral biasing.' I will go on to expound the details of this moral biasing view, which we can draw from Greene's work, and what prescriptive claims he makes in light of this.

3.1.Moral Biasing

Moral biasing, analogously to the wider systemic biasing discussed previously, is a worry that our system one is sensitive to features of situations that are not simply different to system two but different in a way which is maladaptive. To use Greene's main example, it cannot possibly be morally relevant that a person is dropped from a bridge by remote control rather than pushed, yet when the question is framed in these terms the number of people willing to kill in the footbridge case more than doubles (2013, 215).

This, it seems, is deeply troubling given the ability of our system one to override what we might think of as being our more considered judgments.

The evidence for this moral biasing seems to be fairly substantive, Greene's own extensive work in the field has, as previously discussed, used a wide variety of experimental methods to bear this point out (Greene 2014, 701–705). Additionally, these findings have been borne out elsewhere, with one of the more notable examples being Singer's drowning child case (Singer 1972, 229–243). The case is one in which a young child is drowning in a shallow pond and it would cost us only the price of a ruined suit or pair of shoes to save them. Clearly, not saving the child would be morally unthinkable to both our reflective and intuitive moral judgments. Yet, as Singer and other researchers note, for the same small sum of money the life of a small child on the other side of the world could be saved; however, people are drastically less likely to donate even small sums of money to charities that could save these children's lives. It seems, then, that mere spatial distance (i.e., spatial distance in the absence of some other morally important factor) bears on our moral judgment.

A similar study compared two cases where you personally witness a humanitarian tragedy in a country and are asked to donate verses a case where your friend is in the country and shows you a video before asking you to donate. The difference here is clearly not a morally important one, yet it appears to affect people's judgment, in that, people who are not imagining being physically present "drastically" (Greene 2014, 769) less likely to donate. Additionally, it appears race and in-group identification has a fairly substantial baring on these choices in some situations (Swann et al 2010, 1176–1183) and this is before we get into the substantial literature on racial basis in jury decisions (Sommers 2007, 171–187).

3.2.Greene's Prescriptive Claims: Changing Norms

Off the back of the above research, Greene then goes onto make his major prescriptive claims. In being mindful of the Is-Ought distinction (Hume 1738, 3.1.1) (the description that states you cannot derive a normative claim from a merely descriptive one) Greene posits, as I did earlier in this paper, that we can motivate these prescriptive claims on the basis of uncontroversial normative beliefs that we already have. To wit, the descriptive claims gain normative force by being parasitic on our common normative belief set. Thus, no is-ought transgression occurs.

Greene's central prescriptive claim is that we ought to default away from, and indeed distrust, our moral intuitions across a range of situations. This notion is most clearly

expressed in what Greene calls the 'no cognitive miracles principle' which he describes thus:

> The No Cognitive Miracles Principle: When we are dealing with unfamiliar* moral problems, we ought to rely less on automatic settings automatic emotional responses and more on manual mode conscious, controlled reasoning , lest we bank on cognitive miracles. (Greene 2014, 715)

To unpack this principle, Greene takes a 'miracle' here to be a situation wherein, given the way in which our moral intuitions evolved and what they are for and sensitive to, it would be miraculous if they lead us to good moral conclusions. To illustrate this first with a non-moral example, imagine if a quantum physicist presented me with evidence for the correctness of a particular claim and I replied by saying 'that seems wrong to me, intuitively'. This seems like a poor response since quantum physics just isn't the sort of thing about which we should expect human beings to have accurate intuitions. The specific argument laid out by Greene as to whether we should expect our intuitions to be accurate is based upon the notion that our intuitions are primarily based upon experience, be it evolutionary (i.e., useful capacities developed in response experiences by our ancestors being passed down to us genetically) cultural (passed down by cultural experience) or of course personal experience.

In light of this, we can start to determine which sorts of moral problem our intuitions will be able to lead us to good decisions and which situations we ought to adapt a more reflective, cognitively engaged approach. The key, according to Greene, is moral problems that arise from recent (in relative terms) and thus unfamiliar developments. Examples of this would be things like climate change, bioethics, public health, global poverty, terrorism, existential risk (i.e., risk to the continued existence of humanity from nuclear annihilation, pandemic, or AI risk), race and gender relations, etc. This claim, that 'unfamiliar' problems are likely to be where out our intuitions lead us astray is supported by work from Cass Sunstine (2005, 531–573) who posits, with Greene, Kahneman, and Tversky, that though our moral intuitions might be useful in everyday situations once they are confronted with what he terms 'exotic' problems, they seem to lead us astray.

One troubling observation a reader might make here, and this will prove critical later, is that the areas where it seems our intuitions might lead us astray are the most high cost, both morally and practically. Although, as Greene himself notes, these problems might well have components which are 'familiar' to us, the broad strokes of them will not be, as these are all problems that are recent in the grand scheme of human culture and

evolution, and since these things inform our intuitions, their absence suggests that our intuitions are not likely to be adaptive in these areas.

Our strategy then, it seems, should be that we ought to systemically distrust our intuitions across the above class of situations and default away from them when making decisions. When we have a conflict between our intuitions about a case and our more reflective system two thoughts, then we should go with our more reflective thoughts.

To preempt a tempting, though wrong-footed line of objection to the above, why don't our intuitions about 'unfamiliar' cases simply grow more accurate over time? If they are indeed grounded in experience, then surely we should expect our intuitions about these cases to grow more accurate, as we have more experience of them and thus not need to default away from them. This line of criticism fails because, as we found in Kahneman's research, certain sorts of problems are unfamiliar in some deep sense that doesn't seem to change with experience. Recall that people who had spent years of their life studying statistics had no more accurate intuitions about them than anyone else. Now, clearly, their ability to make reasoned, deliberative choices about statistics would be vastly better than average; however, this doesn't seem to effect their intuitions. Similarly, as will be discussed later, people who consider moral problems for a living appear to be subject to the same biases in their moral intuitions as everyone else. We can explain this disparity in terms of the inflexibility of system one judgments, as their speed and efficiency derives from their insensitivity to a broad range of information, which makes them very hard to alter. Thus, though prima faci appealing, this line of argument does not offer a solution to the problem.

3.3. Greene's Prescriptive Claims: The Specifics

Aside from its broad normative implications, what specific impact would Greene's prescriptions have? By which I mean, how specifically should we change our behaviour in response to this account? It seems that, when faced with a choice in areas where we shouldn't expect our intuitions to be useful, we should opt for our more deliberative reasoning instead of our intuitions. We should, as Socrates advises, 'Follow the argument where it leads' (Plato 1966) and go with the conclusion that our deliberative reasoning leads us to, despite intuitive resistance. Additionally, the program that Greene is suggesting implies that we should be more cautious if we only have intuitions about a topic. Say that you're confronted with a question about some complex issue of public health policy, you might know nothing about the topic but have a fairly strong intuition

that a certain policy would be best. Under this account, the prescription appears to be that we ought to remain agnostic.

Additionally, Greene's account, despite appearing, at first glance, somewhat pessimistic about the ability of human beings to make moral choices, has some extremely useful recommendations for how to improve our discourse on matters of ethics, politics, and public policy (Greene 2013, 295–298). He preempts an obvious criticism that simply saying 'think harder about tricky problems' is noble but perhaps useless advice; after all, we seem to already think that we have educated, well-founded opinions on complex problems when, in reality, these are principally intuition-driven. He retorts to this by pointing to a body of research by Fernbach, Rogers, Craig, Fox and Solomon (2013), which appears to show that people can be lead to change strongly held stances on issues of politics and policy by being asked to lay out, in detail, the problem or their solution to it. On discovering that they don't understand the problem, or their solution isn't as well thought out as they thought (or indeed at all), they either abandon their position or lower the credence they have in it.

4. Gigerenzer: Fast And Frugal Heuristics

Having laid out one half of the debate, I will now go to expound the major opposing perspective, namely Gerd Gigerenzer's 'Adaptive toolbox' account. Gigerenzer's 'Adaptive toolbox' account of cognition is at once similar and extremely different to the Heuristics And Biases account. As under Greene, Kahneman, and Tversky's model, Gigerenzer gives an account of cognition wherein heuristics have a central role in our decision making, and indeed gives a very similar descriptive account of heuristics themselves (i.e., that they ignore much of the available information, they lead us to fast decisions, they are in some sense automatic etc). However crucially, unlike the Heuristics And Biases model, Gigerenzer takes our intuitive cognition to be largely adaptive (Gigerenzer and Brighton 2009, 107–143). By this, I mean that he doesn't, broadly, conceive of our intuitions as being systemically biased and indeed holds that they can in fact make better decisions than conscious deliberation in many cases. In the following section, I will go on to expound Gigerenzer's thesis, the key concepts on which he relies, and how his normative claims interacts with those of the Heuristics And Biases Account.

Similarly to the previous account, Gigerenzer defines heuristics as being:

> A strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods. (Gigerenzer and Gaissmaier 2011, 454)

Gigerenzer himself frequently returns to the example of the Gaze Heuristic (Gigerenzer, and Brighton 2009, 108) to illustrate the above, this being a capacity to track moving objects (i.e., baseballs) through the air and to adjust our speed and direction to end up at the spot where it will land. He notes that the baseball playing academic, who attempts to work out where the ball will end up 'manually' will invariably fail since the capacity of any person to do the required calculations using system two is vastly slower, and less accurate, than the heuristic will be. With a single piece of information, namely 'keep your eye on the ball', the heuristic allows us to constantly complete the complex task of catching fast moving objects in a way that feels to us almost effortless.

Interestingly, this example crystallises precisely why and how Gigerenzer takes heuristics to be an adaptive decision procedure, principally through the notion of Ecological rationality.

4.1.Ecological Rationality

I will now go on to carefully expound what the notion of Ecological rationality means in this context, since it is critical to Gigerenzer's account and thus to this debate more broadly. In its most simple form, the idea of Ecological Rationality is that decisions have to be thought of as an interwoven function of mind and environment and cannot be meaningfully talked about as being rational or irrational outside of the environment in which they are taken. Thus, rules which suggest things like, 'for any decision to be rational it must conform to rules X,Y,Z,' and are invariant across situations, are meaningless under this account. The notion is neatly captured by the idea of Simon's Scissors, which Simon coined by positing that:

> Human rational behaviour is shaped by a scissors whose blades are the structure of task environments and the computational capabilities of the actor. (Simon 1990, 7)

The ideas here seem simple enough: human behaviour is grounded inexorably in the situation in which they must behave and we have a set of evolved cognitive capacities that produce behavioural tendencies when combined with the incentive structures created by environments. However, the implications of this account appear to be surprisingly radical.

If we take the account of ecological rationality seriously, we find that certain counter intuitive notions, like Gigerenzer's claim that heuristics that ignore information, even if there would be no cost to acquiring it, can make better choices than a system with greater information, become perfectly rational. Once a definition of rationality becomes

tied to a synthesis of mind and environment, it makes sense that actions that appear to violate rules of choice making (i.e., by not taking into account all available information) can be rational since the incentive structure created by the environment makes them rational. Additionally, this can explain how the moral adaptiveness or otherwise of an intuition can be explained in terms of environment. One of Gigerenzer's main examples is the 'status quo heuristic,' under which people will tend to do what everyone else is doing and ignore other (seemingly very important) information that also exists. The twin examples of how its goodness or badness depends on the environment are relayed presently. His first example (1) is a true example, wherein a group of ordinary polish policemen during the second world war partook in the brutal massacre of Jewish civilians from their own nation (Gigerenzer 2008). The men were given the opportunity to step forward if they did not want to participate, and only a dozen of the 500 men present chose to not to partake. Gigerenzer goes to great lengths to point out that these men were mostly older, ordinary police men, not hardened members of the SS, and that there is good historical evidence that they were not particularly anti-Semitic. His point here is that these men were motivated by the status quo heuristic to not break ranks, and that this heuristics was powerful enough to override their intuition that murdering civilians is wrong.

His second example (2) notes that, in Britain and America, where organ donor laws are opt in (meaning you have to register to be one yourself), rates of individuals registered for organ donation are 17% and 28%, respectively. By contrast, in France and Hungary, where the laws are opt out (meaning you are automatically registered and must deliberately chose not to be) donor rates are 99% and 99%, respectively. His point is that each of these cases involve the status quo heuristic, yet, whilst in (1) it led to egregious acts of horror in (2) it leads to tens of thousands more lives being saved every year. The difference is the incentives unique to each environment, as the heuristics themselves are morally neutral.

4.2. Gigerenzer's Prescriptive Claims: Environmental Design

The major Prescriptive claim that Gigerenzer's account produces, and with Greene he does so by combining interesting empirical facts with uninteresting normative notions, is that moral behaviour is best improved by focusing on how environments and the incentives they give rise to can be better designed to produce the kinds of behaviours we want. This can include simple acts such as minor alterations to the framing of situations (e.g., the officer in change of the policemen asking anyone who *did* feel able to carry

out the task to step forward, rather than singling out those who did not) (Gigerenzer 2008, 6). It can also include higher level policy decisions such as making laws regarding organ donation opt in or, as proponents of nudge theory suggest, things like reducing the size of glasses to reduce excessive drinking, making salads rather than fries a default to encourage healthy eating, etc. Though these last examples are more geared towards public health, it seems that something like this mechanism is what Gigerenzer is advocating. However, it is critical to note that, whilst Gigerenzer is an advocate of something like nudging, broadly defined, he is opposed to its justification on the basis of human irrationality and 'libertarian paternalism (Gigerenzer 2015, 361–383). To wit, he views the standard justification for nudging as buying into the systemic biases account of human cognition that he explicitly rejects.

It is important to clarify here that, whilst Gigerenzer rejects major elements of the nudging program, he is also explicitly advocating for environmental engineering as a way to promote adaptive behaviours (Gigerenzer 2008, 5–6; Gigerenzer 2010, 542). Just as the talented runner must have paths and tracks to run on, in order to make use of her natural gift, so too does Gigerenzer seem to suggest that our intuitive moral thinking will lead us to good choices, if only the environment allows it. To wit, for Gigerenzer, engineering environments is more a matter of allowing people to flourish than leading them because their judgment cannot be trusted. I want to make this point very clear, both because I wish to accurately convey Gigerenzer's stance and because, in light of this, there are certain criticisms of the nudging program that will apply to him and others that won't.

Before I move on to my major argument against Gigerenzer, it is interesting to note how both theories converge on the inflexibility of intuitions. Even Gigerenzer, whose account is deeply sympathetic towards our intuitions, gives prescriptions for how they could be improved by changes to incentive structures and decision environments. Both accounts recognise that we cannot simply will our intuitions to function differently to how they do. Under the descriptions of both accounts, this makes sense for the same reasons. Gigerenzer makes much of how useful it can be that our heuristics are informationally frugal; however, it seems, as has been stated, that it is this resistance to the majority of information that leads them to be so inflexible. In short, both accounts find it necessary to try and account for the inflexible nature of our intuitions.

5. The Central Conflict: Changing Norms Vs Changing Environments

Cashing out the major prescriptive claims of each theory allows us to see what the central conflict between them is. On one hand, we have Greene, who suggests that we

need to change our norms and decision procedures, and, on the other, Gigerenzer, who suggests that we need to change the structures of our decision environments to let our intuitions function better. In this section, I will lay out the central arguments we can bring to bear between the two positions over this key issue and assess them.

I will begin by noting that there is some very trivial sense in which Greene can concede that, yes, our intuitions are only maladaptive given certain decision environments and problems. If we lived in a possible world where we never faced any 'unfamiliar' problems, then, naturally, our intuitions would be unproblematic. So, in one sense, the ecological rationality point is true, but in a very broad, uninteresting sense. The interesting question, upon which the conclusion of this essay will turn, is going to be which of these prescriptions is more plausible in our world, as is.

There is an interesting quote from Sunstine who, when describing the confidence we have in our intuitions regarding 'exotic' cases, says the following:

> They might not deserve to be so firm, simply because they have been wrenched out of the real-world context, which is where they need to be to make sense. (2005, 541)

This quote seems to agree with Gigerenzer's point regarding ecological rationality, yet uses it to support an argument of the kind that Greene is making. Although, as I've suggested, we don't need to look to non-'real-world' examples to find situations where our intuitions won't help us. Now, while this quote seems to make a point that Gigerenzer would agree with (i.e., that our intuitions can become maladaptive if they're removed from an environment with the correct features), I hold that the broader point of the quote is highly damaging to his case. Consider that, if the only situations in which our moral intuitions are useful are ones which are not morally important, then it appears that Gigerenzer's case is reduced to triviality. To clarify more formally why this argument is so damaging:

1. Gigerenzer's view of his own program is that people like Greene et al. are inaccurately characterising our moral choice making as systemically flawed and that, in reality, our moral intuitions can be adaptive in helping us achieve our own moral ends. This is because, under the adaptive toolbox account, our heuristics are capable of quickly and accurately responding to the incentives of the environment. Indeed, he makes much of the fact that the adaptive

toolbox account can be prescriptive. Clearly, the ability for a theory to be meaningfully prescriptive is an important theoretical virtue.

2. However, as Sunstine and Greene's research seems to suggest, the situations in which this occurs are in everyday, morally trivial cases and not in important, morally high-stakes cases.

3. Thus, we can agree with Gigerenzer up to a certain point, but, at the same time, reduce his claims to relative triviality, since, when we talk about 'moral decisions,' we are normally talking about precisely the kinds of situations in which our intuitions fail us. Ergo, our intuitions are not 'morally' useful, in any interesting sense of the word.

To clarify 2 and 3 further, as has been previously stated, our moral intuitions can be useful in familiar decision environments. Aversion to personal harm, defaulting to the status quo etc. seem to be clearly useful tools of maintaining peace and stability, this much the skeptic of intuitions can concede. However, the range of dilemmas in which our intuitions fail us are precisely that, dilemmas. Our intuitions may hold us back from punching someone who we dislike, or move us to comfort a distressed child, but whether or not we ought to do these things isn't really up for question, as we don't agonise over them and they don't appear to be captured by what we mean when we talk about 'moral' choices. By contrast, the exotic dilemmas which our intuitions fail to help us with generally are the sorts of things that the phrase 'moral choice' gets at. Should we give up civil liberties to protect against terror? Could we ever justifiably carry out a permeative strike nuclear strike? If I would save a child dying right in front of me for a small sum of money what else does that commit me to? Whilst the domain of familiar problems, in which Gigerenzer's prescriptions are useful, doubtless contains some moral problems it seems that the majority of them (and certainly the vast majority of high-stakes problems) exist in the domain of the unfamiliar, where Gigerenzer's prescriptions are not useful. Thus, if this reasoning holds, Gigerenzer's prescriptive case is trivial when it comes to solving moral problems.

5.1. Gigerenzer's First Retort

A reply to the above, on behalf of Gigerenzer, would be that, yes, it might be that morally high-stakes decision environments cause problems for our intuitions as it currently stands, but it is a leap in logic to then claim that Greene's solution is the correct one. Given the effectiveness of our intuitions in situations where they do work, it seems that it would be better to try and change these decision environments rather than change our norms. This is true for both a positive and negative reason. The positive reason is that, as Gigerenzer's research seems to bear out, once the correct environment is discovered or created, our intuitions function with startling ease and accuracy – sometimes more so that conscious deliberation. Consider, for example, trying to engage people's system two to convince them to become organ donors (which appears to fail) to simply changing the laws to opt out (which seems to succeed). If such a workaround can be found for other high-stakes moral situations, and, contrary to the above argument, organ donation is an extremely high stakes moral situation, then the model becomes decidedly non-trivial. The negative reason to support the adaptive toolbox account over its competitor is that, as Gigerenzer notes, we can point to fairly damning flaws in cost-benefit thinking when it comes to moral problems (Gigerenzer 2008, 20–23), in that, in an uncertain and complex world, working out a solution on the basis of computational reasoning will prove far too complicated to be useful.

5.2. A Reply to Gigerenzer's Retort: Deep Features of Environments

To reply to the negative point first, whilst the above would be a damaging point if the only options were a heuristic approach or a purely computational one, this is not the state of affairs we find ourselves in. Something like rule consequentialism occupies a middle ground between these two positions. By this, I don't mean a commitment to rule Consequentialism per se but the kind of decision procedure that it implies (i.e., a deliberative process of establishing a rule for a range of situations and defaulting to it). A critic might, at this point, quibble with the degree to which this is really a departure from Gigerenzer's position. I posit that the departure is significant, though subtle. Whilst this sort of rule based decision procedure is perhaps superficially similar to a heuristic based decision procedure (since heuristics are rules), it is a rule established and shaped through deliberative reasoning, which we have access to, rather than by the various unconscious aspects of our cognition, which we don't.

The second element of his case requires more detail to address. This, fundamentally, is the major tension at play between the two accounts (i.e., the issue of ecological

rationality). At base, if Gigerenzer's strategy of environmental engineering and institutional design are not plausible for the problems with which we're concerned, then this entire prescriptive case is implausible. Something that I would suggest makes Gigerenzer's strategy implausible is the notion that decision environments have 'deep features,' for which there is no plausible method of engineering that is workable or non-coercive.

It strikes me that all the concrete examples that Gigerenzer presents of environmental design being successful, or cases where they could be successful, revolve around seemingly surface level features of environments (Bennis et al 2012). Changing the organ donor law to opt-out, though undeniably very effective, involves a relatively simple change in the law. By contrast, people failing to give to charity because of proximity related biases seems trickier. After all, the feature of the environment which causes the problem (i.e., that the people in question are vast distances away) is not something which can be engineered away. Despite the regular presence of advertisements from various charities showing us distressing images of plight, the fraction of people who give to charity, and the amount they give, remains tiny (McKenzie and Pharoah 2011). Thus, it seems that simply trying to use advertising to tug at peoples intuitions isn't working as a strategy. This is more than just a one-off case, for the vast majority of people don't take steps (such as voting for particular policies, etc.) to reduce climate change, despite a wealth of evidence of its effects being regularly displayed on the news that most people consume. We might attribute this too to a kind of proximity bias, albeit a temporal one, insofar as being in full possession of the facts does nothing to effect our intuitions. What ties these, and the other high-stakes cases together, is that their unfamiliarity (and thus propensity to fall victim to our systemic biases) is tied to features of the situation that it would be either impossible or at least far from optimal to remove.

Now, naturally, there are other measures that could be employed to get people to comply with their stated preferences. In the case of global poverty, we could simply make it a default that companies must donate a portion of their employees salary to charity, though, of course, they could opt-out if they wished. This method, however, begins to seem more like an example of coercion than simply allowing people to flourish. This more direct path becomes even more murky when it comes to matters of public policy (note that many of the high-stakes examples concern things like terror policy, immigration, nuclear weapons, etc.) and, as such, there doesn't seem to be a way to influence choice architecture that doesn't also favour one political party undemocratically. As we noted previously, Gigerenzer has already explicitly rejected the more 'heavy handed' approaches to nudging favoured by people like Thaler and Sunstine.

A significant additional problem, which this account falls victim to, is precisely the criticism that Gigerenzer, I believe falsely, levels against accounts like Greene's, namely over complexity. Consider that, for this approach to work, it would need to be the case that for all high-stakes moral problems, a certain account of environmental engineering would need to take place. This, in turn, would require a detailed understanding of the ins and outs of what informs our heuristics on an extending broad and diverse range of issues. This kind of problem seems to be intrinsic to any account that would seek to change the world around us rather than attempt to change us.

By contrast, Greene's decision procedure seems far more flexible, insofar as when we are required to act or form an option on some moral issue, and an intuition presents itself, we need only ask questions which appear to have obvious answers. Does it concern an issue which is, in the grand scheme of history, new or unfamiliar to us? If so, we ought to assume that our intuitive judgment isn't likely to be accurate and default away from it.

5.3. Gigerenzer's Second Retort: Greene's Account Is Useless for Most People

A different line of argument that a defender of Gigerenzer might take would be to posit that Greene's account applies only to cases which the majority of people don't need to deal with. This functions as the inverse of my argument that Gigerenzer's account fails if it only applies in mundane cases, since they might suggest that, by the opposite token, a theory that only applies in extreme cases has equal claim to being trivial. Specifically, they might suggest that the moral sphere of the average agent (i.e., the range of moral problems that they need to engage with) includes things like 'should I drunk drive' or ' should I cheat on my spouse,' etc. Cases of the sort that Greene's theory better applies to are issues that fall within the moral sphere of policy makers and a narrow range of experts. Thus, the critic might suggest that Greene's account represents a decision procedure for a very specific kind of agent making a specific kind of decision. Indeed, Greene himself admits that our moral intuitions work for most people most of the time. The theoretical motivation here might be that if both theories are liable to the accusation of being trivial, then it can hardly be argued that Greene's account is superior to Gigerenzer's in virtue of that criticism.

The above argument is certainly concerning; however, I would suggest it fails, insofar as it incorrectly characterises the moral sphere of the average ethical agent. By this I mean that, even if we think that it matters more that certain people have more carefully considered moral positions on high-stakes issues than others (i.e., policy makers), we can still put forward a convincing case that all moral agents ought to have more careful

attitudes in high-stakes cases. This is because there are at least two significant ways in which the average moral agent can impact these high-stakes issues. Firstly, there are the obvious, 'hard' impacts an agent can have (i.e., voting for a given policy or giving their money to a given charity), and here we might point to the fact that there is significant public devision on issues where there seems to be an extremely high expert consensus (Johnston and Ballard 2016, 443–456) such as immigration, which appears to be an unfamiliar moral issue. The second way in which the average moral agent can effect these issues is the 'soft' way (i.e., by contributing to a general culture of what sort of solutions to problems are acceptable and what behaviours we allow). In general, there are a variety of things we can all do to move the Overton window (the name given to the hypothetical range of acceptable policy suggestions and opinions. This is where the work of Kahneman becomes relevant again. Kahneman notes that the act of gossip can be successful when it comes to shaping peoples attitudes, since being able to predict how your action would be gossiped about, and gossiping about the acts of others, can lead to more adaptive behaviours (Kahneman 2013, 406–409), because we are highly attuned to the faults of others and to how we might imagine others see our faults. My point here is that, whilst the average moral agent is unlikely to ever be in the position of deciding the exact nature of, for example, bioethics policy, personally, they each contribute in both soft and hard ways to the climate which dictates what the policy will be.

5.4. A Third Criticism of Greene: Impossible Prescription

The final criticism that might be levelled against Greene's account is the one that I take to be the most problematic for it. The critic might concede the previous points, that Greene's account is preferable to Gigerenzer's *insofar as it's possible* but might counter than Greene's prescriptions simply aren't possible. For Greene's account to work, agents would need to be capable of defaulting away from their intuitions in unfamiliar decision making. The critic might suggest that we simply don't have that kind of cognitive control, that our intuitions will always be more appealing to us than our more reflective judgments. The weight of this criticism is added to by the fact that advocates of the heuristics and biases view, such as Kahneman and Sunstine, are themselves skeptical of how effectively people can be 'debiased'. Additionally, there is research from, among others, Kushman et al, which suggests that professional philosophers do not seem to be any less susceptible to biases than anyone else, even in examples with which they are familiar (Schwitzgebel and Cushman 2012). Thus, people whose job it is to think reflectively about moral issues still seem to be unable to resist their intuitions. In light

of this, it might be suggested that, even if it is suboptimal when compared to Greene, something like the ecological rationality view is the only game in town.

I would suggest that the above line of criticism, though forceful and intuitive, is mistaken. The first, and most obvious retort, is that we have compelling reasons to reject the empirical claim upon which the criticism rests. As mentioned previously, it seems that when leading people to engage in a certain sort of deliberative process (i.e., thinking through the problem and how their solution solves it), people seem capable of overcoming their intuitive positions. Indeed, Greene's own research seems to show that people with certain sorts of training – the specific case he appeals to is public health officials – seem to adopt a more deliberative decision procedure. We see this evidenced in their willingness to take courses of action, both in their area and in more general cases, that our intuitions would ordinarily reject (Greene 2013, 128–131). The criticism, moreover, suffers from a deeper problem. I would suggest that it falsely conflates skepticism about the current viability of Greene's project with skepticism about his project being worth pursuing. Whether Greene's prescription that we should default away from our intuitions in certain cases fails because it isn't currently possible (and as I have shown this can be questioned), it doesn't follow that we shouldn't try to act with this goal in mind, even if we are currently liable to fail. In short, in addition to the criticism being empirically questionable, it strikes me as unnecessarily defeatist.

6. Conclusion

To conclude, both of these accounts not only provide an invaluable insight into human cognitive architecture but also raise normative concerns that we ought to take seriously. However, what I have sought to demonstrate is that Gigerenzer's account, however successful in the non-moral domain, fails to provide an acceptable defence of our intuitions. It fails, insofar as it becomes trivial when applied to morally important cases, and the concept upon which it depends to resist this triviality, ecological rationality, is a sub-optimal and indeed implausible approach to these problems. Additionally, I have sought to show that Greene's account is useful, plausible, and capable of resisting various criticisms to the contrary. Thus, I conclude that Greene's account is superior and we should indeed distrust our moral intuitions in many significant cases.

Bibliography

Bennis, Will M., Konstantinos V. Katsikopoulos, Daniel G. Goldstein, Anja Dieckmann, and Nathan Berg. 2012. "Designed to Fit Minds: Institutions and Ecological

Rationality" In Ecological Rationality: Intelligence In The World, edited by Peter M. Todd and Gerd Gigerenzer. New York: Oxford University Press.

Fernbach, Philip M., Todd Rogers, Craig R. Fox, and Steven A. Sloman. 2013. "Political Extremism Is Supported by an Illusion of Understanding." *Psychological Science* 24 (6): 939–46.

Foot, Phillipa. 1967. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5: 5–15.

Gigerenzer, Gerd. 2015. "On the Supposed Evidence for Libertarian Paternalism." *Review of Philosophy and Psychology* 6 (3): 361–383.

Gigerenzer, Gerd. 2010. "Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality." *Topics in Cognitive Science* 2 (3): 528–554.

Gigerenzer, Gerd. 2008. "Moral Intuition = Fast and Frugal Heuristics?" In *Moral Psychology, Vol 2, The Cognitive Science Of Morality: Intuition And Diversity*, edited by Walter Sinnott- Armstrong, 1–26. Cambridge: MIT Press.

Gigerenzer, Gerd, and Henry Brighton. 2009. "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1 (1) 107–143.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. "Heuristic Decision Making." *Annual Review Of Psychology* 62 (1): 451–482.

Greene, Joshua D. 2014. "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics." *Ethics* 124 (4): 695–726.

Greene, Joshua D. 2013. *Moral Tribes: Emotion, Reason, And The Gap Between Us And Them*. New York: Penguin Books.

Greene, Joshua D., Fiery A. Cushman, Lisa E. Stewart, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2009. "Pushing moral buttons: The interaction between personal force and intention in moral judgment." *Cognition* 111 (3): 364–371.

Greene, Joshua D., LE Nystrom, AD Engell, JM Darley, and JD Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44 (2): 389–400.

Hume, David. 1738. *Treatise On Human Nature*. New York: Oxford University Press.

Johnston, Christopher D., and Andrew O. Ballard. 2016. "Economists and Public Opinion: Expert Consensus and Economic Policy Judgments." *The Journal of Politics* 78 (2): 443–456.

Kahneman, Daniel. 2013. *Thinking Fast And Slow*. New York: Farrar, Straus and Giroux.

Kahneman, Daniel. 2003. "A Perspective on Judgment and Choice: Mapping Bounded Rationality." *American Psychologist* 58 (9): 697–720.

McKenzie, Tom, and Cathy Pharoah. 2011. "How generous is the UK? Charitable giving in the context of household spending." *CGAP Briefing Note 7*. London: CGAP.

Plato. 1966. *Plato in Twelve Volumes, Vol. 1*. Translated by Harold North Fowler. Cambridge: Harvard University Press.

Schwitzgebel, E., and F. Cushman. 2012. "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." *Mind & Language* 27: 135–153.

Simon, Herbert. 1990. "Invariants Of Human Behaviour." Annual Review of Psychology 41: 1–19.

Singer, Peter. 1972. "Famine, Affluence And Morality." *Philosophy And Public Affairs* 1 (3): 229–243.

Sommers, Samuel R. 2007. "Race And The Decision Making Of Juries." *Legal And Criminological Psychology* 12 (2): 171–187.

Sunstine, Cass. R. 2005. "Moral Heuristics." *Behavioral and Brain Sciences* 28 (4): 531–573.

Swann, WB, A Gómez, JF Dovidio, S Hart, and J Jetten. 2010. "Dying and Killing for One's Group: Identity Fusion Moderates Response to Intergroup Versions of the Trolley Problem." *Psychological Science* 21 (8): 1176–1183.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing Of Decisions And The Psychology Of Choice." *Science, New Series* 211 (4481): 453–458.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science, New Series* 185 (4157): 1124–1131.

Journal of Cognition and Neuroethics

Review of *Moral Brains: The Neuroscience of Morality*

James William Lincoln
The University of Kentucky

Biography

James William Lincoln is a PhD candidate in Philosophy at the University of Kentucky. His primary research project, broadly speaking, focuses on moral perception and its role in making reliable moral judgments. Presently, his research uses a comparative philosophical approach in an effort to articulate a theory which accounts for an agent's ability to see salient moral properties in everyday life. He argues that our perceptual faculties are, at least in the moral domain, cognitively penetrable by our beliefs and attitudes. This means that moral features or properties of the world are perceivable and that one must possess the appropriate moral beliefs and affective/emotional attitudes regarding the contents of morality if that perceptional experience is to be trusted as justification for an accurate moral judgment of any present situation. His work utilizes Feminist Affect Theory, Marcusean Social Theory, Buddhist Moral Psychology, the Philosophy of Perception, and Moral Neuroscience to unpack this topic because he believes that each of these domains offers conceptual tools from which to understand the moral subject as a member of the larger social environment and as an integrated cognitive system.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2018. Volume 5, Issue 2.

Citation

Lincoln, James William. 2018. "Review of *Moral Brains: The Neuroscience of Morality*." *Journal of Cognition and Neuroethics* 5 (2): 75–81.

Review of *Moral Brains: The Neuroscience of Morality*

James William Lincoln

Abstract

S. Matthew Liao's recent publication of *Moral Brains: The Neuroscience of Morality* represents a valuable contribution to the field of moral neuroscience. In this review, I provide a brief summary of Liao's collected anthology of essays by philosophers and scientists that explore the intersection of neuroscience and ethical theory. I claim that this text is a excellent resource for philosophers and scientists alike and briefly argue for a cautious engagement with its contents because of empirical limitations commonly associated with philosophical investigations into refining the object of study.

Keywords

Review, Moral Neuroscience, Ethics, Cognitive Science

Review of

Liao, S. Matthew. 2016. *Moral Brains: The Neuroscience of Morality*. New York: Oxford University Press.

Review

By creating a collaborative space for neuroscience and ethical theory, the field of moral neuroscience seems poised to provide invaluable insights into our moral lives. *Moral Brains: The Neuroscience of Morality* is an accessible and instructive contribution to this field. In its editor's own words, this collection "is the first to take stock of fifteen years of research" (Liao 2016, 33). Its arrival onto the scene as the "first" to do this is, however, less important in my assessment than what the volume attempts to accomplish and its addition of thirteen original works to field. As editor, S. Matthew Liao seems intent on providing a guide from which to introduce the uninitiated to almost two decades of work regarding the intersection of neuroscience and ethics. Something that, prior to this volume's publication, has been virtually impossible to find.

Moreover, this collection appears to be a genuine attempt to foster a collaborative conversation between the neuroscientific and philosophical communities. Unfortunately, professional philosophy has a recent history of resistance to the inclusion of empirical data into its methodology. However, this volume represents a substantial effort among scientists and philosophers to survey moral neuroscience's major issues. Moreover, it

does this while maintaining a willingness to engage questions regarding the value or admissibility of neuroscience findings to ethical theory. Julia Driver, Jesse Prinz, James Woodward, Joshua Greene, S. Matthew Liao, and many of the other contributors represented in this collection have been vanguards for this kind of interdisciplinary scholarship. This volume is an effective invitation into a field which asks us to acknowledge that ethical theory should be sensitive (while not necessarily assenting) to theories about cognitive mental structures. As such, Liao's collection attains part of its value from the fact that it successfully puts established and emerging scientists and philosophers into meaningful conversation with each other.

Moral Brains gives its reader an introductory picture of the landscape one might encounter while exploring the larger body of scholarship in moral neuroscience. To that end, this volume is organized into four parts and includes an invaluable introduction by Liao. His introduction gives a helpful overview of the research responsible for inspiring the field by reviewing several landmark studies during the 1990s. It also briefly discusses the debate regarding the admissibility of neuroscience data to ethical theory while introducing the reader to the major topics explored throughout the rest of the volume. These topics include such things as motivational internalism, the role of emotions and reasoning in moral judgments, moral intuitions, and the intersection of neuroscience and normative ethics. Overall, Liao's introduction accomplishes a difficult task. It provides the philosopher with access to the science, the neuroscientist with a general idea of the philosophy, and the completely uninitiated with tools to find footholds for further engaging the subject.

Part one, titled "Emotions vs. Reasons," tackles the issue of sentimentalism and rationalism in moral decision-making. Prinz's argument for a sentimentalist theory of moral judgment in "Sentimentalism and the Moral Brain" is appropriately followed by Kennett and Gerrans's argument, in "The Rationalist Delusion?: A Post Hoc Investigation." Prinz argues that psychological evidence supports a sentimentalist view of moral judgment even though uncertainty plagues the neuroscientific research on this point. Prinz's view maintains that emotions, traditionally understood in the history of western philosophy as passions, are the driving force behind moral judgments (66–69). In contrast, Kennett and Gerrans respond to this kind of view by pointing out how deliberative reflection and reasoning over time is essential to making moral judgments (77). They, thereby, present a rationalist counterview in opposition to Prinz's kind of sentimentalism. They argue, essentially, that moral deliberation's relationship with reason and diachronic agency is more important than Prinz, or those that might hold similar views, would want to admit (83). The section concludes with Woodward's piece on

emotion and cognition which argues that the very distinction between emotion and cognition in cases of moral judgment is a dubious dichotomy (88-89). Moreover, he observes that if this rigid delineation between emotion and cognition is problematic, then there exists a questionable assumption in both rationalist and sentimentalist positions (113).

Part two, titled "Deontology verses Consequentialism," gives the reader a general sense of how moral neuroscience approaches issues regarding moral intuitions and their role in moral judgments. This section begins with a reprint of Greene's 2014 article, "Beyond Point-and-Shoot Morality," in which he observes that deontological forms of moral deliberation utilize emotional thereby making them automatic forms of judgment formation whereas consequentialist forms are shown to be more grounded in rational deliberation. Taking the neuroscience to support a dual-process view of judgment formation, wherein emotions and rationality simultaneously yet independently shape one's judgments, Greene argues for a kind of epistemic caution in regards to deontological moral claims. He believes such judgments are unreliable because of their dependence on automatic rather than deliberative judgment formation processes (130–134). Julia Driver's "The Limits of the Dual-Process View" responds to Greene's claim by arguing that his concerns only seem to apply to a narrow set of intuitionist moral views and that more complicated theories of moral judgment avoid his critical gaze (157). Stephen Darwall, in "Getting Moral Wrongness into the Picture," argues that there are forms of rule consequentialism that would be "characteristically deontological" in the sense that Greene is concerned. Darwall thereby suggests that the kind of recklessness associated with deontological judgments also seem to apply to consequentialist judgments as well (168). Thus, Darwall claims, there is no ground to privilege consequentialist over deontological judgments epistemically. The section concludes with a reply to Darwall and Driver by Greene. It is, however, unclear if Greene is successful in his response.

Part three, titled "New Methods in Moral Neuroscience," turns the discussion towards issues of cognitive functioning and the selections within attempt to use neuroscientific observations to argue for the presence of moral predispositions in our neurological structures. Blair, Hwang, White, and Meffert observe that emotion-learning systems contribute to full moral development by shaping norm expectations and that, from a neurological perspective, there are four kinds of norms associated with this growth. These include disgust-based, harm-based, and justice-based norms as well as norms prescribed by social convention (195). Oliveira-Souza, Zahn, and Moll attempt to flush out the neurological foundations of moral cognition by applying a lesion study to brain-damaged patients using neuroimaging techniques (203). Crockett uses serotonin studies

to explore its impact on moral judgment and behavior. She argues that moral judgment decisions are closely related to one's neuromodulator levels and stress even in cases where serotonin's presence had no detectable impact on the subject's mood (239). Borg examines the nature of unjustified violence and suggests that rodent models of negative Intersubjectivity have the potential to effectively develop treatments for clinically violent patients (267). Ultimately, the hope here, as I understand it, is that such research would help us understand, in a more robust sense, our mental relationship to morally relevant actions.

The final section, titled "Philosophical Lessons," explores the implications of moral neuroscience on normative ethical claims. Kahane's contribution argues for three main things. First, that "there are multiple ways to validly draw potentially interesting normative conclusions from empirical premises" (301). Second, "that findings about the internal structure of our moral psychology, or about its underlying neurobiology, will have only a limited role to play in such arguments" (301). And lastly, that if we want neuroscience to contribute to ethical theory, then we cannot let these fields operate independently of the other on these issues (301). He also claims that we might need to rethink our approach to empirical research as a consequence of the observations mentioned above. Liao's contribution to this section argues that intuitions are not heuristics and that one consequence of this insight is that Greene's view that deontological intuitions tend to be inaccurate because of their automatic (i.e., heuristic) nature is unsupported (328). The section concludes with a piece by Sinnott-Armstrong which argues that morality is not unified. It observes of "judgments that are intended to be about morality … [that they are] are not unified by any single common and distinctive feature that enables important generalizations about distinctive properties of those judgments" (335). He goes on to suggest that "scientists should isolate smaller classes of judgments" by content and context, rather than employing a top-down method, which begins by making the distinction between moral and non-moral judgments. I take Sinnott-Armstrong to be suggesting that an alternative methodology, which he calls bottom-up, shifts us towards taxonomic rigor by accepting that we cannot "study morality all at once" (350).

In general, I believe this collection is a valuable contribution to the field of moral neuroscience because it gives its reader access to a new perspective on three important questions in ethical theory, questions Liao discusses in his introduction. These include "How do moral judgments differ from non-moral judgments?", "Are moral judgments based on or driven by reasons or emotions?", and "To what extent can moral judgments be reliable?". The importance of these questions, I hope, is clear from what I have written

thus far, but we might also observe that neuroscience and ethical theory still operate independently of each other in both their methodology and pre-theoretical assumptions. The history of western philosophy since Descartes is one in which a belief in the rational subject has become a kind of ideology. This ideological predisposition to conceive of the subject as a rational being often manifests in the presumption that our rational and emotional systems are distinctly isolated or that our emotive existence consistently corrupts our capacities to make moral judgments. Additionally, the theoretical preference for reason over emotion motivates a social convention in many domains of professional philosophy to acknowledge that the moral agent is capable of compartmentalizing or unifying moral judgment under the faculty of reason. Moreover, the neuroscientific community can, from its first step into the metaphorical lab, carry pre-theoretical beliefs about morality or moral judgments which simultaneously limit the scope of the salient questions and the methods for their investigation. Sinnott-Armstrong's piece hints at this observation, and it is refreshing to see similar claims from some of the volume's other contributors who advocate for a reimagining of the scientific approach to studying the moral features of the brain and the philosophical approach to the moral agent.

To conclude, the strengths of this volume are numerous. It is designed for an academic audience while being accessible to the non-academic hobbyist with minimal difficulty. For those looking to take the first step into moral neuroscience scholarship, you would be hard-pressed to find something as valuable as this collection. However, there are some limitations to the volume that the potential reader should be aware. First, it is limited in its scope in virtue of its status as an anthological collection of essays. The reader will need to spend time investigating the studies mentioned by Liao during the introduction to grasp the full history of the field because such additions would seemingly have been cumbersome to include in this kind of text. Second, one needs to be aware that debates on things like motivational internalism or cases of psychopathy involve numerous disputes about language. Debates in motivational internalism vs. externalism, for example, often result in interlocutors talking past each other or holding different criterion for saying that an agent possesses a moral belief. Philosophy's engagement with psychopathy cases also seems to center on what it means when we say so-and-so understands moral reasons. In these instances, the philosophy attempts to refine its sense of the object of study rather than its understanding of the object under study, and empirical projects are better suited to the latter kind of efforts. In these cases, what it means to have a moral belief, what it means to make a judgment, and so on are somewhat isolated from any aid by empirical science.

Overall, one should engage this text cautiously aware of these limitations. Until such time that philosophers and scientists have a richer collaborative history which includes, not just the investigations of ethical questions, but the construction of those questions, it is best to keep an eye on the distinction between investigations into refining the object of study and empirical research into the targeted object under study. What Liao has provided in the publication of this volume is a start to that history and a model for furthering an invaluable interdisciplinary relationship between cognitive science and philosophy's investigation of ethical theory.

cognethic.org