Self-Awareness of Culpability: The Mainspring of Moral Behavior?

By

Mukesh Eswaran*

Department of Economics
University of British Columbia

March 2010, Revised February 2011

ABSTRACT

We propose that the experimental findings informing the recent arguments for a Universal Moral Grammar can be explained through the concept of awareness of culpability of the self in an evolutionary context. We argue that the doctrine of double effect, which is usually invoked to explain such findings as those in trolley experiments, is really a derivative principle. By incorporating the tacit but crucial role played by the awareness of culpability of the self of the decision-maker in these dilemmas, we are able to understand why this doctrine carries force. In addition, we provide an argument to show how and why moral codes can differ across societies. One conclusion our argument points to is that moral behavior is the result of a special feature of the general-purpose evolutionary vehicle of self, namely, its proclivity to minimize its culpability. Finally, by explaining some puzzling anomalies that arise in the context of the doctrine of double effect, our analysis arrives at a theory of the origin of morality: the evolutionary imperative to temper self-orientation.

Key Words: morality, universal moral grammar, self, culpability, natural selection, doctrine of double effect

Acknowledgements: I would like to thank Murat Aydede, Charles Blackorby, David Donaldson, Vinayak Eswaran, Patrick Francois, Dominic Lopes, Ara Norenzayan, Margaret Schabas, and the participants of the Departmental Colloquium in Philosophy at the University of British Columbia for valuable comments on an earlier draft (which was entitled "Universal Moral Grammar or the Self?"). All shortcomings, however, are my own.

* Mailing Address: #997-1873 East Mall, Vancouver, B.C., CANADA V6T 1Z1
Email: eswaran@econ.ubc.ca

I. Introduction

Some scholars have argued that our brain is hardwired for a universal grammar of morality, along lines analogous to the universal grammar of language.[1] This view has gathered force recently and has found comprehensive expression in *Moral Minds: The Nature of Right and Wrong* of Marc Hauser and in writings of John Mikhail.[2] Hauser, Mikhail and others who espouse this view draw their inspiration from Noam Chomsky's work, which posited some five decades ago that, gauging from the rapidity with which children learn complicated rules of grammar, it must be the case that humans come equipped with a universal grammar for language.[3] In this view, different languages are merely applications of this universal grammar. Each language is characterized by some universal principles, each of which has a limited number of options ("parameters") that have to be set in the learner's mind. (For example, does the preposition come in the beginning of the sentence or in the end?) Different languages may have different parameters for these principles. Once a child manages to set these parameters for the specific language she is exposed to, the universal grammar she innately possesses enables her to learn the language with great rapidity and precision. This occurs despite the fact that there is inadequate stimulation ("poverty of stimulus", in Chomsky's telling phrase) from the environment. Furthermore, children do not exhibit conscious awareness of the rules that guide their grammatically correct speech. By drawing an analogy with the notion of Universal Grammar, Hauser, Mikhail, and others argue that humans are hardwired with a universal "moral organ" which enables them to assess moral dilemmas.[4]

To make the case for a Universal Moral Grammar Hauser (2006) presents results,

---

[1] E.g. Harman, G. (2000), "Moral Philosophy and Linguistics", Ch. 13 *of Explaining Value and Other Essays in Moral Philosophy*, Oxford University Press, New York; Dwyer, S. J (1999), "Moral Competence", in *Philosophy and Linguistics*, Murasugi, K. and R. Stainton (eds.), Westview Press, Boulder, pp. 169-190.

[2] Hauser, M. (2006), *Moral Minds: The Nature of Right and Wrong*, Ecco, New York; Mikhail, J. (2007), "Universal Moral Grammar: Theory, Evidence and the Future", *TRENDS in Cognitive Science*, 11, pp. 143-152.

[3] Chomsky's early books on this subject were *Syntactic Structure*, The Hague, Mouton, 1957, and *Aspects of the Theory of Syntax*, MIT Press, Cambridge, 1965.

[4] For a critical review arguing that Universal Moral Grammar lacks a grammatical structure, see Dupoux, E. and P. Jacob (2007), "Universal Moral Grammar: A Critical Appraisal", *TRENDS in Cognitive Sciences*, 11, pp. 373-378.

reported in Hauser et al (2007)[5], on the experiments conducted with thousands of people, eliciting their judgments in moral dilemmas. These results reveal that people with diverse social, religious, and ethnic backgrounds have similar responses. They tend to make their assessments with great rapidity, but they cannot often articulate the principles that led to their assessment. Mikhail (2007), too, bases his arguments for a universal moral grammar on the basis of such experiments. In making their claims, these authors heavily rely on the doctrine of double effect. According to this doctrine, in order to achieve a highly desired good it is permissible to inflict unintended (but foreseen) harm, but it is not permissible to inflict harm that is intended.

In this paper, we argue that the experimental findings cited above can be explained by recognizing and incorporating the role of a special feature of self in our theorizing. By 'self', we shall mean that conscious entity which oversees the perceived wellbeing of the psycho-physical organism that an individual refers to as 'I'; everything else it construes as 'other'. We claim that in evaluating moral dilemmas the self of the decision-maker (or evaluator) inevitably and instinctively intervenes and that her judgments are colored by this intercession. We propose that, by examining how the attributes of self got shaped in an evolutionary environment—where violence may have been expedient but not inevitable in some circumstances—we may discern how a moral code would have arisen that is consistent with the findings cited by Hauser (2006), Hauser et al (2007), and Mikhail (2007). In other words, our paper provides a theory of the origins of the morality embodied in these experimental findings. In effect, we argue in this paper that the doctrine of double effect that is usually invoked to explain behavior in many moral dilemmas is really a *derivative* principle: it derives from the role played by an aspect of self in an evolutionary setting. The approach we adopt, therefore, is a positive one in which we seek to explain the empirical geneses of moral principles.[6]

In the sort of moral dilemmas that have led analysts to posit a Universal Moral Grammar, the evaluators assessing the morality of others' actions are themselves uninvolved. This might suggest that evaluators' self-interest could not be instrumental in

[5] Hauser et al (2007), "A Dissociation Between Moral Judgments and Justifications", *Mind & Language*, 22, pp. 1-21.

[6] For a normative analysis of the latter sort, see Quinn, W.S. (1989), "Actions, Intentions, and Consequences: The Doctrine of Double Effect", *Philosophy and Public Affairs*, 18, No. 4, 334-351.

shaping their choices. Furthermore, evaluators often pass different verdicts on the morality of actions in different scenarios that have identical consequences. Thus, it might appear that these judgments divulge their 'true' or 'absolute' moral values. We argue that this is not so. Self colors all our moral judgments because self is a crucial, if tacit, actor in these moral decisions. The self-interest of the evaluator *is* involved. Thus we cannot correctly identify the principles guiding moral judgments while ignoring this fact (as in arguments that invoke the doctrine of double effect, which is an axiomatic principle). When we do account for the role of a specific feature of self (a feature which we subsequently dub 'self-awareness of culpability'), we can explain not only the choices that are consistent with the doctrine of double effect but also some choices that, while consistent with our intuition, appear anomalous in the light of the doctrine.

There is evidence from neuroscience that the self is implicated in moral dilemmas and so the premise of this paper is not without empirical foundation. Functional magnetic resonance imaging (*fMRI*) studies have sought to identify which areas of the brain are involved in moral, as opposed to non-moral, judgments. A recent study has found that dilemmas requiring intentional harm to others more heavily called on the brain circuitry responsible for emotions than those that required no such harm.[7] An earlier study found that, among the many brain regions that process self-relevant stimuli, one particular region that represents emotions was the main one involved in self-referencing.[8] The authors claim that this area is responsible for processing personally relevant emotional stimuli. Yet another *fMRI* study has revealed that moral sensitivity to matters pertaining to justice and care implicated three regions of the brain, all of which are involved in self-referential processing.[9]

We argue here—and this is the fundamental claim of this paper—that the underlying principle that guides choices in all moral dilemmas is the objective of *minimizing self's culpability*, where culpability accrues to actions that, in an evolutionary

[7] Borg, J.S. et al (2006), "Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation", *Journal of Cognitive Neuroscience*, 18, pp. 803–817. Similar results are also reported in Greene, J.D. et al (2004), "An fMRI Investigation of Emotional Engagement in Moral Judgment", *Science*, 293, pp. 2105-2108. See also other references cited later.

[8] Fossati, P. et al (2003), "In Search of the Emotional Self: An fMRI Study Using Positive and Negative Emotional Words", *American Journal of Psychiatry*, 160, pp. 1938-1945.

[9] Robertson, D. et al (2007), "The Neural Processing of Moral Sensitivity to Issues of Justice and Care", *Neuropsychologia*, 45, pp. 755-766.

setting, would have proved damaging to survival prospects. To put it differently, moral decisions are driven by the goal of moderating the *self-orientation* of one's actions because such moderation, paradoxically, would have maximized reproductive fitness under the given conditions in our evolutionary past. By self-orientation we mean the tendency to exclusively focus on our own survival to the detriment of that of others. In effect, we claim that morality arises from a need to blur the self/other distinction to some degree, that is, to draw a less trenchant distinction between ourselves and others.

There are some moral values (e.g. the prohibition of murder) that are widespread in that they are common to most societies and cultures. Other moral values, however, may differ greatly across cultures and societies. For example some societies may find infanticide acceptable but others may deem it abhorrent. We suggest a mechanism showing how this is possible, using the framework we propose here. In this view the explanatory work is done by the perceived culpability of self, which is the overarching governor of an individual's survival mechanism. If some aspects of morality appear to be universal, it is because self and, more specifically, its sense of culpability is universal.

The theory proposed in this paper is sympathetic to the view recently espoused by Prinz (2008, 2009), who has argued against the need for positing a Universal Moral Grammar.[10] In his view, morality arises in a cultural context where it is enforced by emotional disapproval as a by-product of mechanisms that evolved for something else. We argue that this mechanism is really a special feature of the self, and that individual morality is just one more institution in the arsenal of self to enhance survival in a social setting.

The rest of the paper is as follows. In the next section, using two moral dilemmas we summarize the argument for innate morality that is made in the literature. In Section III, we review the evolutionary origins of self, introduce the notion of self-awareness of culpability and identify the role it plays in the making of moral decisions. We provide the evolutionary underpinnings of the decisions that are attributed to Universal Moral Grammar. In Section IV we show how culture can impinge on morality. This illustrates

[10] Prinz, J.J. (2008), "Resisting the Linguistic Analogy: A Comment on Hauser, Young, and Cushman", in *Moral Psychology*, Vol. 2, Sinnott-Armstrong, W. (ed), MIT Press, Cambridge, MA; Prinz, J.J. (2009), "Against Moral Nativism", in *Stephen Stich and His Critics*, Murphy, D. and M. Bishop (eds.), Wiley-Blackwell Press, Malden, MA.

how and why moral codes can differ, despite the fact that the basic faculty responsible for making moral decisions is universal. In Section V, we illustrate how the role we attribute to self-awareness of culpability enables us to rationalize our intuition in some anomalous moral dilemmas and demonstrate that the origin of morality stems from the evolutionary benefits of tempering the claims of self by softening the self/other dichotomy. We present some concluding thoughts in Section VI.

II. The Experimental Basis for the Claims of Innate Morality

The case for a Universal Moral Grammar is based on the premise that, since there is a poverty of stimulus with regard to moral knowledge (as with linguistics), moral knowledge must be innate [Dwyer (1999), Harman (2000), Mikhail (2007), Hauser et al (2007)]. In order to communicate the flavor of the empirical basis of the theory, in this section we briefly summarize two moral dilemmas that have received attention from researchers. We shall describe two that Hauser (2006) and Hauser et al (2007) use in their experimental settings. Mikhail (2007) also refers to such experiments in his recent arguments.

Denise's Dilemma: An out-of control trolley is going down a track on which five hikers are walking. The hikers cannot jump off the tracks and all five of them will certainly die if the trolley continues. There is a sidetrack on to which the trolley can be diverted by throwing a switch. But there is one hiker walking on this alternative track and he will be killed as a result of this diversion. Denise has the option of flipping the switch to divert the trolley on to this alternative track. Should she flip the switch, thereby killing only one hiker instead of the five who were on the trolley's original track?

Frank's Dilemma: The situation here is similar to the one above, but there is no sidetrack onto which the trolley can be diverted. Frank is watching from an overbridge as an out-of-control trolley in front of him is heading for five hikers walking on the track behind him. Standing right next to Frank is a very obese man. If Frank pushes the man off the overbridge, he will fall on the tracks and will stop the trolley. The obese man is

sure to die, as a result. Should Frank push the man on to the track, thereby killing one person but saving five others?

As reported by Hauser et al (2007), most people (90%) say "Yes" to the quandary facing Denise and most (90%) say "No" to the one facing Frank. This is irrespective of the decision-makers' race, gender, age, religion, nationality, ethnicity, etc. There is no difference in the outcomes if Denise and Frank both opt to save the five hikers at the expense of one individual. But people pass different moral judgments on their actions. The harm brought about by Frank's and Denise's actions have been referred to as 'direct' and 'indirect' harm, respectively, where the former is part of the action while the latter is the result of some other action.[11] We believe that it is not so much direct and indirect harm that is the issue as much as it is intended or unintended harm. The puzzle then is why people find intended harm more objectionable than unintended harm.

To explain the findings on the sort of dilemmas facing Denise and Frank, Hauser (2006) and Mikhail (2007) (and some others before them) argue that people are endowed with an innate moral sense. This moral sense forbids them to do harm to achieve some greater good, but permits them to do harm if it is an unintended (though foreseen) consequence of wanting to do good. St. Thomas Aquinas, who first discussed the morality of doing harm to achieve good, called this the *doctrine of double effect*.[12] This doctrine, we note, makes an assertion without providing a reason for it. Since this principle is usually not taught to children, it is unclear how they learn to use it. This is why Harman (2000), Hauser (2006), and Mikhail (2007), among others, deem it to be part of a Universal Moral Grammar that humans putatively are innately endowed with.

Mikhail (2007) posits that, in making moral evaluations, humans behave as lawyers governed by two beliefs: intentional battery is forbidden, and the principle of double effect. The latter we have explained above; the former he describes as the prohibition of willful harm to another without consent. Using these two deontic principles as axioms, he demonstrates that he can provide an explanation for the evaluations in the

[11] Royzman, E.B. and J. Baron (2002), "The Preference for Indirect Harm**",** *Social Justice Research*, 15, pp. 165-184.
[12] For the pioneering contemporary consideration of this, see Foot, P. (1967), "The Problem of Abortion and the Doctrine of the Double Effect", *Oxford Review*, 5, pp. 5-15.

sort of moral dilemmas posed above. One difficulty with such an axiomatic approach, however, is that it leaves unexplained *how* the principles invoked as axioms become salient to moral decision-making.

Nichols (2005) is skeptical of the innateness of morality; he argues that there is a plausible case for rule innateness (rule comprehension) but not so for moral nativity because he is unconvinced that the doctrine of double effect is the best explanation for moral intuition.[13] He attempts to offer an explanation of the doctrine of double effect by positing that, when humans engage in moral decision-making, they employ two independent systems: a deontological one (which absolutely forbids some actions, like murder) and a utilitarian one (which assesses the costs and benefits of actions). The former first screens the actions and, if they get passed the screen, the latter assesses them. If the actions pass both tests, they are deemed to be permissible; if they fail even one, they are deemed impermissible. The difficulty with this explanation is that it suffers from the same drawback as Mikhail's (2007) theory: the deontological system remains unexplained. The evolutionary theory we offer in this paper resolves this issue.

Greene and Haidt, in independent and joint work with various co-authors, offer a different explanation.[14] Backed by neuroscientific evidence, they argue that emotions are undoubtedly involved in moral decisions. They draw a distinction between 'personal' harm, in which harm is done to a particular person and which is not the result of a deflection of a threat to someone else (as in Frank's case), and 'impersonal' harm which is not personal (as in Denise's case). They view this distinction as being crucial to explaining moral behavior, since the former entails greater involvement of the emotions. Experimental results of brain scans while participants were answering questions in non-moral and moral dilemmas (the latter of personal and impersonal kinds) revealed that emotions were engaged in moral dilemmas. The more personal these dilemmas were (as in Frank's case), the greater was the extent to which emotions were implicated. Hauser et

[13] Nichols, S. (2005), "Innateness and Moral Psychology", in *The Innate Mind: Structure and Contents*, (eds.) Carruthers, P. et al., Oxford University Press, New York, pp. 353-369.
[14] Greene, J. and J. Haidt (2002), "How (and Where) Does Moral Judgment Work?", *TRENDS in Cognitive Sciences*, 6, pp. 517–523; Greene, J.D. et al. (2001) "An fMRI Investigation of Emotional Engagement in Moral Judgment", *Science*, 293, pp. 2105–2108; Greene, J. (2005) "Cognitive Neuroscience and the Structure of the Moral Mind", in *The Innate Mind: Structure and Contents*, (eds.) Carruthers, P. et al., Oxford University Press, pp. 338–352; Haidt, J. (2001), "The Emotional Dog and its Rational Tail: a Social Intuitionist Approach to Moral Judgment", *Psychological Review*, 108, pp. 814–834.

al (2007), though, argue that results from other experiments (which we do not discuss here due to space restrictions) suggest that the personal/impersonal distinction cannot be the whole story. Greene and Haidt papers cite evolutionary arguments in support of their hypothesis about the salience of emotions in personal harm. While they are agnostic about moral nativity in the strong sense that normative rules are hardwired, they acknowledge that the nature of the cognitive processes that are instrumental in implementing moral behavior depends on evolutionary design and that moral beliefs themselves probably do have a genetic component [see Greene (2005)]. While this strand of the literature is relevant to this paper because they touch on evolutionary arguments for moral behavior, the specific evolution-based theory we develop in the next section differs considerably from these because it offers a fairly detailed account of *why* evolution may have had a hand in shaping moral behavior.

Having seen the sort of scenarios that have led to the proposal that there is a Universal Moral Grammar, in the next section we propose a theory that explains the experimental findings. We shall argue that we need nothing more than recognize the role played by a specific aspect of self in order to understand behavior in the above dilemmas. This role generates outcomes *as if* there were a Universal Moral Grammar.

III. Self-Awareness of Culpability and its Tacit Role in Moral Dilemmas

In this section we argue that the doctrine of double effect really derives from a more primitive principle. We claim that we can identify the empirical basis of this doctrine and, furthermore, pinpoint the general principle that generates behavior consistent with it. It is our contention that such moral rules are nontrivially influenced by how the self of the evaluator is affected when making moral judgments.

The Evolutionary Emergence of Self

A unique feature of humans is self-awareness. This sense of self-awareness—the notion that one is an entity separate from the environment and even from other conspecifics—has many gradations. The important advance in human evolution from the

point of view that is pertinent here is that, at some point, humans developed the cognitive abilities to conceive of themselves as individuals. The self is responsible for a person's agency or the executive functions based on volition. This is one important aspect in which humans are different from other species, where stimulus-response drives much of behavior. The self retains memories of past experiences, considers alternative scenarios for the present, plans and plot strategies for the future, etc.[15] Sedikides and Skowronski (1997) refer to this as the 'symbolic self', and they argue that the emergence of hunting and of social relations in the *homo erectus* stage of hominid evolution was especially important in the emergence of this self.[16] More generally, it was Gallup (1982) who first argued that self-awareness evolved because it conferred survival advantages.[17] Povinelli and Cant (1995) have put forward the intriguing hypothesis that the self evolved in large-bodied arboreal apes.[18]

We should note that although evolution has shaped the individual self in humans, there is also a substantial social component to this entity because humans evolved in social groups. There is considerable evidence to suggest that the group to which a person belongs influences her thoughts, preferences, and behavior, not only consciously but also unconsciously.[19] The strong feeling of an individual self promoted by natural selection is tempered by social considerations that impinge on survival chances when there is competition between groups. Such inter-group competition is conducive to the emergence of in-group/out-group distinctions in one's preferences.[20] This view is consistent with the claims of self-categorization theory in psychology, which views self as an entity that sees

---

[15] For a good overview of the subject, see e.g. Kihlstrom, J.F., J.S. Beer, and S.B. Klein (2003), "Self and Identity as Memory", in *Handbook of Self and Identity*, (eds) Leary, M.R. and J.P. Tangney, Guilford Press, New York. The article by Baumeister, R.F. and K.D. Vohs, "Self-Regulation and the Executive Function of the Self" in the same volume is also excellent.

[16] Sedikides, C. and J.J. Skowronski (1997), "The Symbolic Self in Evolutionary Context", *Personality and Social Psychology Review*, 1, pp. 80-102. They also provide an extensive survey of the literature relevant to the evolution of self.

[17] Gallup, G.G. (1982), "Self-awareness and the Emergence of Mind in Primates", *American Journal of Primatology*, 2, pp. 237-248.

[18] Povinelli, D.J. and J.G.H. Cant (1995), "Arboreal Clambering and the Evolution of Self-Conception", *The Quarterly Review of Biology*, 70, pp. 393-421.

[19] For a review of the extensive evidence and literature, see e.g. Devos, T. and M.R. Banaji (2003), "Implicit Self and Identity", *Annals of the New York Academy of Sciences*, 1001, pp. 177-211.

[20] For an evolutionary model of how in-group and out-group preferences could have been shaped in an evolutionary setting, see, Eaton, B.C., M. Eswaran, and R. Oxoby (2007), "Us and Them: Evolutionary Origins of Identity and Some Economic Consequences", forthcoming, the *Canadian Journal of Economics*.

itself as “I” and “me” or as “we” and “us”, depending on whether the social context emphasizes inter-person differences or inter-group differences.[21]

The importance of hunting arose in human evolution when *homo erectus* moved from the forests (where *homo habilis* is believed to have resided) to the savannas. Sedikides and Skowronski (1997) argue that this entailed a change from scavenging (and foraging) to hunting and would have put pressure on the cognitive abilities of hunters. Improvements in cognition that enabled humans to follow fast-moving animals, assimilate and assess rapidly changing signals, plot strategies to increase success in hunting, etc. would have been favored by natural selection. Furthermore, team work in hunting—especially big game hunting—required a level of cooperation between humans that would have been absent in earlier stages of evolution. This would have led to social interactions and the emergence of norms to counter free-riding behavior. In this view, the activity of hunting and the social structure it engendered constituted important ingredients that went into the shaping of a sophisticated sense of self that emerged in *homo erectus* during the late Pleistocene. We note, however, that it is not hunting alone that could have led to these developments and so our argument here does not hinge on the role of hunting. Any kind of large-scale team activity (such as defense of the tribe) would have generated selective pressures ultimately leading to similar refinements of the self.

Leary and Buttermore (2003) have offered an alternative, perhaps richer, story of the evolution of self, drawing on the view posited by Neisser (1997) that there are five different aspects to self.[22] Inferring the aspect of self that we would be required to produce the various pieces of archaeological evidence that is available at hand, Leary and Buttermore speculate on the time at which it would have evolved. In their view, the symbolic self emerged at around the time that the great “cultural revolution” occurred (40, 000 to 60, 000 years before the present), in which technology, painting, language, music, etc. emerged in sharp contrast to what existed before that. It is their contention that these cultural activities require a self-concept, which strikes us as very plausible.

---

[21] See e.g. Turner, J.C. et al (1994), “Self and Collective: Cognition and Social Context”, *Personality and Social Psychology Bulletin*, 20, pp. 454-463.

[22] Leary, M.R. and N.R. Buttermore (2003), “The Evolution of the Human Self: Tracing the Natural History of Self-Awareness”, *Journal for the Theory of Social behavior*, 33, pp. 365-404; Neisser, U. (1997), “The Roots of Self-Knowledge: Perceiving Self, It, and Thou”, in J.G. Snodgrass and R.L. Thompson (eds.), *The Self Across Psychology*, New York Academy of Sciences, New York, pp. 19-33.

For *homo sapiens sapiens*, protection of the self is the predominant principle that governs the survival of the individual. We may, for our convenience here, think of the self as being composed of various aspects: the mental self, the emotional self, and the physical self. The mental self would constitute a person's mental representation of who she believes she is, her identity. The emotional self would constitute the emotions that she utilizes (anger, envy, empathy, love, etc.) to sustain her sense of self. These emotions may also accompany various autobiographical memories that are lodged in her brain. The physical self would stand for the fundamental identification of her consciousness with her body as 'me'. These three components of the self would be mediated, respectively, by phylogenetically the most recent parts of the brain to the oldest, that is, they are listed here in reverse order of their evolutionary emergence.

There is increasing evidence from neuroscience on the brain mechanisms that are responsible for various aspects of the self. There is now evidence that self-awareness is processed in the right hemisphere of the brain.[23] It is now understood how the feeling of ownership of the body arises through the processing of simultaneous visual, tactile, and positional inputs received by the brain.[24] One of the quintessential features of the self is the sense of agency, the feeling of being the originator of actions and, therefore, as being responsible for them.[25] Neuroscientific evidence is mapping the brain areas responsible for agency.[26] The brain circuitry that is responsible for feelings of empathy is being brought to light.[27]

Explaining Behavior in Moral Dilemmas

We are now ready to embark on our explanation for why people behave in the way they do in the sort of moral dilemmas considered by Hauser (2006), Hauser et al

[23] Platek, S.M. et al (2004), "Where am I? The Neurological Correlates of Self and Other", *Cognitive Brain Research*, 19, pp. 114– 122.
[24] Ehrsson, H.H., Spence, C., and Passingham, R.E. (2004), "That's My Hand! Activity in Premotor Cortex Reflects Feeling of Ownershipof a Limb", *Science*, 305, pp. 875-877.
[25] Gallagher, S. (2000), "Philosophical Conceptions of the Self: Implications for Cognitive Science", *Trends in Cognitive Sciences*, 4, pp. 14–21. For a review of neurobiological evidence on agency, see Blakemore, S.-J., and C. Frith (2003), "Self-Awareness and Action", *Current Opinion in Neurobiology*, 13, pp. 219-224.
[26] Farrell et al (2003), "Modulating the Experience of Agency: a Positron Emission Tomography Study", *NeuroImage*, 18, pp. 324–333.

(2007), Mikhail (2007), and others. Self is affected by feelings of *culpability* triggered by contemplated actions that are harmful to others. Self-awareness of culpability is an emotion that is felt privately but has its roots in social relationships. In scenarios where communal feeling is warranted, if there has been a transgression that has hurt one or more members of the community the transgressor usually feels culpable.[28] Communal feeling probably first arose in evolution in the context of kin selection but later expanded to other members of the community on whom survival depended, even if less directly. Self-awareness of culpability requires some degree of empathy; we have to possess the faculty by which we can put ourselves in the shoes of another in order to determine whether we are culpable for inflicting harm on them. The feeling of empathy, which has been identified by some as the basis of morality, has evolutionary underpinnings. The selective advantage in evolution conferred via feelings of culpability is not hard to identify. Transgressions that were hurtful would lead to retribution because of the instinct for self-preservation. An unpleasant emotion like self-awareness of culpability would have helped prevent actions that would have called for retributive actions with deleterious survival consequences for the original perpetrator. This would have conferred an advantage on people who, through experience, anticipated the feelings of culpability accompanying actions harmful to others. (And if such actions were undertaken, self-awareness of culpability would have led them to normalize relations by making amends.)

We prefer to invoke the concept of 'self-awareness of culpability' in our explanations as opposed to the more common emotion of guilt. The latter, too, has evolutionary origins, as has been argued for example by Joyce (2006).[29] The difference is that self-awareness of culpability is more innate and less easily manipulated by cultural factors and idiosyncratic incidents than is its cousin, guilt. For example, a Catholic who has not gone to confession for a few months may feel guilt, but this is a result of upbringing and socialization; there is nothing evolutionary in this emotion here.

---

[27] Jackson, P.L., A.N. Meltzoff, J. Decety (2005), "How do we Perceive the Pain of Others? A Window into the Neural Processes Involved in Empathy, *Neuroimage*, 24, pp.771-779.

[28] An excellent survey of the theoretical and empirical literature on guilt is provided by Baumeister, R.F., A.M. Stillwell, and T.F. Heatherton (1994), "Guilt: An Interpersonal Approach", *Psychological Bulletin*, 115, pp. 243-267. See also Tangney, J.P. and R.L. Dearing (2002), *Shame and Guilt*, Guilford Press, New York.

[29] Joyce, R. *The Evolution of Morality*, MIT Press, Cambridge, Massachusetts, 2006

Furthermore, culpability is a notion that is usually invoked when there is harm or untoward consequences to others; there is no such requirement for guilt.

There is another distinction between guilt and self-awareness of culpability. The position taken in this paper is that self-awareness of culpability is an emotion; it may invoke the thought process, but not necessarily so. It was argued by Darwin (1879, Ch. IV) that conscience requires some degree of intellectualizing, so that guilt requires thought.[30] Joyce (2006, Ch. 2), citing Darwin, goes much further and posits that guilt is not possible in the absence of language. We claim that self-awareness of culpability is an emotion that does not necessarily require thought and, therefore, can operate even without the use of language. Self-awareness of culpability may well have emerged before the appearance of language. This would explain why participants in trolley experiments are often observed to arrive swiftly at judgments but are unable to articulate reasons for their decisions.

It should be noted that the sense of culpability clearly requires, as a precondition, the sense of self. But the existence of self (or belief in it) is only necessary, not sufficient; much more is required. For a person to feel culpable, she has to have an entity within her that takes responsibility for her actions; there has to be a sense of agency already in place. That the self induces a sense of agency in humans is beyond dispute; it is, in fact, one of the defining characteristics of self.

The sense of culpability is invariably the result of harm contemplated or done to *others*; it does not accompany harm done to oneself. This is because self, being the overarching custodian of the body, has been evolved by nature to promote the body's survival; the instinct for self-preservation is very strong. Except in pathological cases, self is not usually given to deliberately doing harm to the body it is identified with. So there was no need for nature to evolve a defense against the remote possibility of intentionally inflicting harm on oneself. If any defense was required, it was against the inveterate tendency of the self to harm others to get what it wants. It is partly to temper the excesses of this tendency—beyond the point where it ceases to aid survival and begins to undermine it—that the sense of culpability has evolved.

[30] Darwin, C. (1879), *The Descent of Man, and the Selection in Relation to Sex*, Penguin, 2004.

Let us briefly reiterate the point here. Hurting someone is permissible if he has harmed us, because evolution has selected retaliatory action as something that is useful for survival. But it is impermissible for us to harm someone who has done us no harm. By harming him, we are inviting retaliatory action (from the victim or his kin) that is damaging to our own survival. We feel culpable when we willfully harm an innocent person. Self-awareness of culpability prevents us, largely (but not entirely) through emotional responses, from engaging in actions that would reduce our chances of survival due to retaliation for needless harm we inflict on others. The sense of culpability is an adaptive emotion: it promotes survival. This reasoning also explains why this emotion accompanies *intended actions*, not unintended ones. If we do unintended harm, we may regret the inadvertent harm done but we would not feel culpable. And for the unintentionally harmed person, there is no survival advantage in punishing these actions: since they were not volitional and perhaps even outside the volitional control of the incidental perpetrator, retaliation has no deterrent effect in this instance.

The difference between intended and unintended harm is why the law draws a distinction between actions that are premeditated and those that are performed without such premeditation, perhaps in passion. To the extent that the law is meant to deter premeditated actions, the penalty for it should be—and is—higher than that for unpremeditated ones. When a person is killed, this distinction is at the heart of the separation between murder and manslaughter.[31]

Prinz (2009) has emphasized the role of guilt in the evolution of morality. The mechanism he suggests, however, is different from the one we are espousing here. He argues that guilt follows from the sadness we feel when we lose the friendship (because of the harm we have inflicted on them) of someone we empathize with. He ties guilt to sadness. In Prinz (2008, p. 78), he articulates his position thus: "Guilt is sadness that has been calibrated to acts that harm people about whom we care."[32] But people may and do feel guilty even when they harm strangers. We argue that the role of self-awareness of culpability was to thwart retaliatory actions that would have reduced the reproductive fitness of the perpetrator. This sense of culpability would prevent harmful actions

[31] Hauser (2006, Ch. 3, p. 148)
[32] Prinz, J. J. (2008), *The Emotional Construction of Morals*, Oxford University Press, New York.

towards all innocent people, whether they are known to us or they are strangers for whom we do not particularly care.

In the experiments we have discussed, participants judge not their own actions but those of others. But how does one's own sense of culpability translate into a moral sense that judges the actions of others? There are many non-exclusive avenues through which this might occur, each of which is supported by evidence from neuroimaging. For one, there is a close connection between the neural networks that identify 'self' and those that identify 'other'. In fact, we would expect this to be so. After all, the 'self' and 'other' are defined as more or less mutually exclusive and collectively exhaustive categories; the very definition of 'self' implicitly defines the 'other. Therefore, it would be very surprising if there were no overlap in the neural networks that bifurcate the perceived world into these two conceptual realms. Considerable overlap in the relevant networks *is* found in neuroimaging studies. The initiation of actions, the simulation of one's actions and those of others, and observations of others' actions all stimulate widely overlapping regions of the brain.[33] There are shared networks that enable individuals to project thoughts and feelings on to others. Indeed, these shared networks are presumed to be the neural bases of theory of mind and of feelings of empathy that characterize an essential distinction between humans and other species.[34] Also of relevance to this question is the theory of common-coding.[35] According to this theory, perceptions of events and intended or associated actions are commonly coded in the brain. So whether one intends an action or perceives the action of another, the same neural networks are activated. In an overview of the current knowledge in neuroscience of self-consciousness and agency, Jeannerod (2003, p. 142) maintains that the pathological condition of compulsive imitation suggests a more general rule: when observing someone else's action, we are never far from performing it ourselves.[36] This is not to suggest, by any means, that the conceptual border between self and other is completely erased by shared networks. Self-

[33] For a brief overview of the literature, see Decety, J. and J. Grezes (2006), "The Power of Simulation: Imagining One's Own and Others' Behavior", *Brain Research*, 1079, pp. 4-14.
[34] Decety, J. and T. Chaminade (2003), "When the Self Represents the Other: A New Cognitive Neuroscience View on Psychological Identification", *Consciousness and Cognition*, 12, pp. 577-596.
[35] Hommel et al (2001), "The Theory of Event Coding: A Framework for Perception and Action Planning", *Behavioral and Brain Sciences*, 24, pp. 849-937.
[36] Jeannerod, M. (2003), "Consciousness of Action and Self-Consciousness: A Cognitive Neuroscience Approach", in *Agency and Self-Awareness*, (eds) J. Roessler and N. Eilan, Clarendon, Oxford, pp. 128-149.

awareness still maintains a distinction between the two concepts. As we might expect to be the case, to facilitate the first-person and third person perspective-taking that we routinely witness in humans, the overlap between the neural networks for 'self' and 'other' is by no means complete.[37] In view of these findings it is entirely plausible that, through various avenues, emotions and thoughts pertaining to self (such as culpability) become salient to judgments on others' actions.

Let us now reconsider the trolley problems we summarized in the previous section. The self-awareness of culpability that has been grafted onto our sense of self enables us to explain why it is that most people would want Denise to throw the switch, killing one person but saving five. Because the death of the one hiker on the adjacent track is unintended, there is no culpability associated with it to the decision-maker. So killing the one person to save five is deemed permissible by our moral code even though the death of that one person is foreseen. But in Frank's case there has to be an intentional action in which Frank has to push the obese man over the bridge. Though the ultimate intention is to save the five hikers, there has to be a deliberate action intended to harm an innocent man. This is ruled out by an inherent sense of culpability that nature has evolved in us. When third parties witness others doing intended harm their morality, informed by the adaptive emotion of their own self-awareness of culpability (for reasons outlined above), gets triggered and delivers a verdict on the action. Hence, killing the one man to save five is deemed impermissible in Frank's case because the evaluators perceive intentional harm. Ultimately, it is the sense of culpability of self in us that is responsible for the intuitive feeling that it permissible for Denise to throw the switch but impermissible for Frank to push the obese man over the bridge.

The higher non-human primates may have a sense of self but, since it is very rudimentary, self-awareness of culpability would not have evolved to the same extent as in humans. Consequently, while non-human primates might exhibit an embryonic sense of morality, we certainly cannot expect a highly nuanced code of ethics to have evolved

[37] Ruby, P. and J. Decety (2003), "What You Believe Versus What You Think They Believe: A Neuroimaging Study of Conceptual Perspective-Taking", *European Journal of Neuroscience*, 17, pp. 2475-2480.

in them.[38] Despite the fact that some behaviors are similar in humans and non-human primates, the mental states that generate the behaviors are very different.[39]

Our argument brings home the point that judgments in moral dilemmas involving third parties inextricably involve the concept of self. An inquiry into the evolution of morality that makes no reference to the sense of self and its culpability misses an essential reason for why moral codes probably evolved in the first place. Since morality obviously has meaning only in the context of a self (one's own or another's) undertaking or evaluating actions, to leave self out of the picture is to leave out an actor who may be the tacit point of reference in these actions or evaluations.

We see that we can understand the experimental findings that Hauser et al (2007) relate by taking account of how the external situation impinges on the culpability of self of the decision-maker. The self is that entity in us that claims authorship of our decisions. In these moral dilemmas, given the clear need to do the right thing, those decisions that generate the least intense feeling of culpability for the author of these actions will be embraced. The self seeks to evade responsibility for intentional harm because the latter generates the feeling that one is culpable, an adaptive emotion. When the consequences to others are the same (one dies, five are saved), the choice is determined by the consequences to self. That is why, we submit, *moral dilemmas with identical consequences to others are treated differently; the consequences to self are different.* This rationale we offer here also provides the evolutionary underpinnings for the doctrine of double effect.

There is some recent evidence from neuroscientific investigations that the self plays a role in judgments in moral dilemmas. In a functional magnetic resonance imaging exercise undertaken when participants were responding to moral quandaries, Borg et al (2006) found that neural circuitry responsible for self were frequently implicated. In their words (p. 808): "Our results suggest that moral judgment may utilize more self-referential processing than even important judgment about one's possessions or

[38] For an argument that there is continuity in the moral capacity of non-human primates and humans, see de Waal, F. (2006), *Primates and Philosophers: How Morality Evolved*, (ed) S. Macedo and J. Ober, Princeton University Press, Princeton, NJ. Also of interest is *Self-Awareness in Animals and Humans*, (eds) S.T. Parker et al, Cambridge University Press, Cambridge, 1994.

[39] For evidence on this, see Povinelli et al (2000), "Toward a Science of Other Minds: Escaping the Argument by Analogy", *Cognitive Science*, 24, pp. 509-541.

livelihood, consistent with the idea that we consider our morals to be crucially defining parts of who we are and who we want to be." These authors, along with those cited earlier (in footnotes 7 and 14), find that when moral dilemmas entail intentional harm, more emotional processing is called into play. As they observe, the evidence is consistent with the fact that participants in their experiments are frequently unable to articulate what led them to declare certain actions immoral. As Fossati et al (2003, p. 1943) note, emotions "generally signal issues related to the self". Casual observation suggests that emotions are more tethered to the psycho-physical entity a person takes herself to be than are thoughts.[40] The enlisting of the brain circuitry that processes emotions in experiments with moral dilemmas is very suggestive of the involvement of the self of the evaluator.

Our hypothesis that moral judgment entails the unavoidable but tacit involvement of the evaluator's self does not imply that moral judgments require conscious reflection. Having been shaped by evolution, the self can undertake actions by recruiting emotions that circumvent thinking altogether. In fact, it is very likely that many moral judgments are arrived at not through the operations of thought at all but derive from a more visceral source.[41] Neuroscience has identified that the 'emotional brain' (comprising the amygdala, the hippocampus, the anterior cingulate cortex, and the hypothalamus) is seriously involved in the making moral decisions.[42] This region of the brain evolved well before the region responsible for our advanced cognitive abilities and responds more automatically than the latter.

Haidt (2001) has persuasively argued that emotions play an important role in moral decision-making, and are not the result of rational thought.[43] He puts forth the view that moral intuition is not based on reflection: "Moral intuition is a kind of cognition, but it is not a kind of reasoning." (p. 814) Rather, in Haidt's view reasoning provides a rationalization after the fact. In our view, emotions are an aspect of self for they are

---

[40] Indeed, even self-transcending emotions like admiration and compassion are found to recruit brain circuitry that processes self-relevant inputs. [See Immordino-Yang et al (2009), "Neural Correlates of Admiration and Compassion", *Proceedings of the National Academy of Sciences*, 106, pp. 8021-8026.]

[41] See Haidt (2001) discussed below and Greene et al (2001) cited earlier.

[42] See, Tancredi, L.R. (2005), *Hardwired Behavior: What Neuroscience Reveals About Morality*, Cambridge University Press, New York, Ch. 4.

[43] Haidt, J. (2001), "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment", *Psychological Review*, 108, pp. 814-834. Also see his recent review, "The New Synthesis in Moral Psychology", *Science*, 316, pp. 998-1002.

usually not experienced as disembodied feelings; they are invariably accompanied by an unstated but acknowledged sense of ownership. When one experiences fear, for example, there is a tacit but definite feeling of ownership that says, "It is I whom am afraid, or I am afraid." (We submit that this sense of ownership, in fact, accompanies thoughts, too.)

Haidt and Joseph (2004) adopt a moderate ground with regard to innateness and emphasize that morality is the outcome of the interaction of a few innate modules in the brain with learning.[44] They identify humans as having innate intuitions about seeing people suffer, about cheating when reciprocity is warranted, and about notions of purity. In their view, moral intuitions stemming from these modules generate socially constructed morals through appropriate moral emotions triggered in an environment that allows for learning.

Prinz (2008, 2009) has taken a strong position against the claim that morality is innate. By innateness he means the presence of modules in the brain that are dedicated to specific purposes. He has shown that, indeed, even the injunctions for not harming innocent people or for respecting authority or against incest—which might be deemed moral universals—are far from universal in a form that might justify any claim to innateness. He has argued that moral norms evolve culturally and are enforced by emotions (such as guilt, embarrassment, disgust, etc.) that evolved for other reasons. Our claim regarding the role of self is consistent with Prinz's view. Self is the general-purpose entity that nature has evolved in humans to oversee the survival of individuals in a social setting and morality, we claim, is an outcome of a specific aspect of it: self-awareness of culpability.

Incorporating the role of self in moral decision-making, as we have done here, explains why intention seems to matter so much in evaluating the morality of actions. A sense of agency stemming from self is certainly present in all willful action. Unwarranted harmful actions are deemed evil and are attributed to self. However, one can conceive of willful inaction as also being evil when action is warranted (harm done by omission). We see, then, that action itself is not the essential focus of moral decisions. Rather, we claim it is *intention.* That is why, we suggest, the principle of double effect—for which Hauser

[44] Haidt, J. and C. Joseph (2004), "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues", *Dædalus*, Fall, pp. 55-66.

et al (2007) find compelling evidence—emphasizes that, to bring about a desirable outcome, the harmful effect (though foreseen) must be unintended for the action to be deemed permissible. It is the *will* that is the crux of the issue. This is presumably why St. Thomas Aquinas, when formulating the principle of double effect, put an emphasis on intention.[45] In our view, the principle of double effect may be seen as an attempt at circumscribing the egocentrism of self through societal pressure that was internalized by self in an evolutionary setting.

Though we claim that the sense of culpability responds to intention, not merely to action, we must acknowledge that the intensity of this emotion is far greater when an evil intention is translated into action. This is very likely because the brain attaches a stronger sense of agency to action than to inaction. As a result, the self would deem itself responsible for evil that it actively brings about; it would hold itself less guilty for evil that it has intended but only passively facilitated through inaction. This would explain the experimental results demonstrating that, *given the intention to do harm*, acts of omission are viewed as less objectionable by subjects than acts of commission. Indeed, subjects are often found to accept greater harm by avoiding action than by undertaking them.[46] Holding constant the intention to do harm, an act of commission carries with it the immediacy of experience that innately confers a sense of responsibility; in an act of omission, the sense of responsibility is an inference, a mere thought, and so carries less force. This is an explanation that rationalizes the "doing and allowing" deontic axiom that is often invoked to explain judgments in moral dilemmas: it requires more justification to do harm than to allow it.

Hauser et al (2007) find evidence that behavior consistent with the doctrine of double effect is not a product of thought, for if that were so people who are educated would have invoked the principle more than the less educated—which the data do not suggest. Because it is the involvement of self-awareness of culpability (a visceral feeling) that is crucial in moral dilemmas in our view, the thought process is inessential. Furthermore, the involvement of self explains why the evidence in Hauser et al (2007) is

[45] St. Thomas Aquinas, *Summa Theologiae: A Concise Translation*, (ed.) T. McDermott, Metheun, 1989, Ch. 11, p. 390.
[46] Spranca, M., E. Minsk, and J. Baron (1991), "Omission and Commission in Judgment and Choice", *Journal of Experimental Social Psychology*, 27, pp. 76-105.

seen to hold irrespective of the gender, age, religion, ethnicity, etc. of the participants. Since self's sense of culpability is a product of evolution, its involvement in moral dilemmas is universal.

Another important point that our view brings out is that none of the moral codes that have evolved through natural selection can claim to be absolute, because self is not absolute. It is universally present in all humans, but it is not absolute. To appreciate this point we only need to recognize that the society and culture we live in seriously impinges on our sense of identity.[47] So even if self's sense of culpability determines moral codes, it is possible for these codes to differ by culture. One could interpret this by saying that the parameters of the putative Universal Moral Grammar would be set at different values in different societies. This point leads us to the next section.

IV. Explaining Cultural Differences in Morality

Would moral codes differ by culture? Hauser (2006) claims that they would, but does not provide us with any indication as to how this might come about. He asserts that we are hardwired to have a moral grammar, but (as with language) the precise parameters can be set to different values for different cultures. The vagueness is unsatisfactory. Unless we specify the mechanisms through which this might happen, almost any sort of moral code we observe can be made to appear consistent with the view that we come equipped with a Universal Moral Grammar.

Prinz (2009), in arguing against moral nativity, has vigorously pushed the idea that culture is largely responsible for morality. Emotions like guilt and embarrassment arose for other reasons but are useful in enforcing moral norms. Embarrassment, he argues, is an emotion that evolved to shun unwarranted attention. Consistent with this argument, and complementing it, we would argue that embarrassment probably has a great deal to do with perceived loss of status, too. Since people with status have greater claim over resources and, therefore, tend to have higher reproductive fitness, natural selection would have favored status-seeking as an adaptive trait. The loss of face

[47] For a review, see e.g. Wood, J.V. and A.E. Wilson (2003), "How Important is Social Comparison?", in *Handbook of Self and Identity*, (eds) Leary, M.R. and J.P. Tangney, Guilford Press, New York.

following any public action in which one falls short of social expectations leads to embarrassment. Aversion to the attendant loss of status would have enabled anticipated embarrassment to drive people to measure up to social expectations in all walks of life, including morality. But for one to feel embarrassment, of course, there has to be a sense of self which an individual takes to be "I" and which feels culpable for her actions.

The arguments we have made in this paper also provide us with a mechanism that can explain how and why moral codes may differ across cultures. This mechanism works through the perceived benefits and costs to self. Since humans are social animals, the penalty for bad behavior has a large social component. Apart from retribution from the injured party, other members of society may ostracize the perpetrator; relatives may disown him, friends may shun him, and so on. In the hunting and gathering societies of the past, punishments such as ostracism would have had very harsh consequences—perhaps even death. So a fear of social sanctions would have evolved. A strong sense of culpability, too, would have evolved to thwart actions that would invite such sanctions.

Consider the fact that in some cultures infanticide is morally acceptable while in others it is deemed abhorrent. In a society living under very harsh ecological conditions, resource scarcity may lead to approval of infanticide. By limiting the number of children in this manner, the probability of survival of the remaining members of a family may be increased. An adult who commits infanticide would not meet with severe disapproval from others in this society because the latter are likely to be in precisely the same predicament. An individual brought up in this society would probably not learn to disapprove of the practice of infanticide; this practice would be compatible with her moral code. As an adult she may herself practice this if the circumstances so warrant.

In another society where resources are abundant, additional children may not impose a huge burden on resources and consequently may have little effect on the survival chances of other members of the family. In fact, the older children may even be of some help in looking after the younger ones. In such a society, infanticide may be looked upon with horror. In the event someone commits infanticide, it would invite considerable societal sanctions. In this society, the self-awareness of culpability associated with killing infants would be so great (even in the absence of legal sanctions) that infanticide would strictly violate its moral code.

This argument shows how moral codes can differ across cultures and yet be consistent with evolved behavior. In a resource-scarce environment, survival of the existing population may be enhanced by infanticide while in a resource-abundant environment infanticide may be not only unnecessary but also detrimental to the maximization of biological fitness. These two societies will evolve different moral codes because of the fear of the penalty individuals would incur in violating the respective codes. The self internalizes the cost of social sanctions in the form of the sense of culpability, and this determines its moral responses.

The theory proposed here for the mechanism of how morality originates also explains why moral behavior is common to all humans—it is because self-awareness of culpability is universal. It is a product of evolution and every human is endowed with it. The existence of this sense of culpability is enough to explain the behavior of humans in moral dilemmas. There is no need to presume that there is a special 'moral organ' in the brain. There is no compelling evidence to suggest that self has special regions in the brain dedicated to self-relevant processing.[48]

We should emphasize that the theory we have proposed here is not merely a more parsimonious alternative to Universal Moral Grammar; it offers more. As we shall see in the following section, it also explains moral judgments that are anomalous from the point of view of the doctrine of double effect. Furthermore, our theory points to the insight that the origins of morality lie in the evolutionary advantage conferred by a moderation of self-orientation to bounds beyond which the latter would have become maladaptive.

V. Some Anomalies in Moral Dilemmas Explained

Explicit recognition of the tacit but crucial role played by self-awareness of culpability in moral decisions allows us to sort out situations in which our moral decisions are not in conformity with what is indicated by the doctrine of double effect. For example, consider the standard example of a soldier who throws himself on a grenade in order to save his comrades. The harm he does himself is intended; he knows that he can save his comrades

only by killing himself. Since the soldier intends to kill himself as a means to an end, the principle of double effect would deem this action morally impermissible. And yet our intuition suggests that not only would such an action be deemed permissible, it would be considered laudable—it is an action that most people might hold up as an example to others. Likewise, the final act of Sydney Carton in *A Tale of Two Cities* would be condemned by few as morally impermissible. In cases like these, the principle of double effect, which Hauser (2006) and Mikhail (2007) espouse as part of the Universal Moral Grammar, leads us astray for it would have us proclaim such actions as impermissible. But our intuition says otherwise.

Anomalies of the above sort arise because the theoretical underpinnings of the doctrine of double effect have not been correctly identified. The essence of the matter in the case of the soldier and of Sydney Carton is *not* whether the harm that is done is intentional or is an unintended side effect. The relevant criterion, as we have argued, is whether the action creates the feeling of culpability for the ego of the evaluator.

Most people (including devout Catholics) would not forbid the action in which a soldier sacrifices himself for his comrades because, as we have already argued, there is no feeling of culpability associated with the harm done to oneself.[49] Likewise, most people see nothing to condemn in Sydney Carton's willful laying down of his life for another. It is one thing to inflict willful harm on others; it is quite another to do so on oneself. Harming others without cause is to give vent to self; this is deemed unethical. Harming oneself to help others is to undermine self; that is deemed ethical. Acknowledging the role self plays in these decisions resolves the anomalies that plague the doctrine of double effect.

We may also apply our reasoning to the issue of suicide. Why is it that people usually have serious misgivings about deeming suicide a permissible action? How does the case of suicide differ fundamentally from the one in which a soldier sacrifices himself to save his comrades? This question has led people to argue that the proportional good that is done as a consequence of the action is an implicit but important part of the

[48] The literature from experimental psychology and cognitive neuroscience on this issue is reviewed by Gillian, S.J. and M.J. Farah (2005), "Is Self Special? A Critical Review of Evidence from Experimental Psychology and Cognitive Neuroscience", *Psychological Review*, 131, pp. 76-97.

doctrine of double effect. It is presumed, in other words, that the magnitude of the good effect that is the desired goal outweighs the unintentional (but foreseen) harm that is done in achieving that goal. We think this is an erroneous claim in this instance. When a person commits suicide, he obviously thinks that the benefit of death to him outweighs the costs to him. And yet most people would intuitively deem suicide impermissible. So proportionality is not what is at stake in such cases.

Suicide is often (but not always) an act of extreme self-absorption. It is frequently an act in which a person abandons all hope and succumbs completely to despair. It is true that, at some level, all unhappiness ultimately stems from preoccupation with self. But the despair that leads to suicide is almost absolute in its absorption. The person is clearly miserable but he is self-absorbed to the extent that he is unable to draw his attention to anything other than his own misery. It is because of this reason, we claim, that most people instinctively deem suicide an impermissible act; suicide is intrinsically *self-oriented* to the extent that the person contemplating it can see nothing outside his own perceived sense of wretchedness. In contrast, a soldier who dies to save his comrades shows concern for others; that act is intrinsically *other-oriented*. In neither case is the sense of culpability necessarily involved for, as we have seen, there is little reason for it to be associated with harm inflicted on oneself.[50] It would appear that *it is the rejection of self-orientation* that constitutes the foundation of moral judgments. It is this principle that informs the moral judgments that people instinctively make. Since self is a product of evolution, this principle is universal.

In some cases, suicide is not a decision made in despair but the result of detached deliberation. This is often the case of terminally ill patients who seek physician-assisted suicide, for example. These are typically people who either cannot perform the act by themselves or cannot find the means to do so. Third party observers of this type of suicide are not likely to be as firmly opposed to it as to suicide in general.[51] The reason, we suggest, is that physician-assisted suicide is not the outcome of self-absorption but,

---

[49] 'Self-awareness of culpability' and 'guilt' can be clearly seen as separate concepts here. Guilt is an emotion religion might instill in the case of suicide, but evolution would not instill the sense of culpability.
[50] Guilt may arise at the thought of suicide in those whose religions forbid suicide, but even people who are not particularly religious deem suicide impermissible.

rather, is a choice made for rational reasons—often out of concern for family members who may face considerable difficulties as a result of, say, the person's prolonged illness.

VI. Conclusions

All good judgments require the evaluator to distance herself from the situation she is assessing and view it impartially. But the very nature of passing a moral judgment puts limits on the extent to which this can be done. It is the premise of our theory that self is inevitably involved in the process, a premise supported by neuroscientific evidence [e.g. Borg et al (2006)]. Evolutionary hardwiring kicks in with the self's sense of culpability. Consequently, moral judgments are inherently *personal* in the sense that they unavoidably involve the self of the decision-maker. This role is tacit, and so it is easily overlooked. By ignoring the role of self-awareness of culpability in moral decisions we may well be overlooking an arguably important determinant of these decisions. This oversight may necessitate the invocation of various principles in an ad hoc manner to rationalize human behavior. In this paper, we have explicitly recognized and analyzed the role that culpability of self plays in moral dilemmas. By doing so, we are able to understand outcomes that have been recently attributed to the presence of a Universal Moral Grammar. We are also able to understand the basis of behavior that, while intuitive, are puzzling from the point of view of received principles such as the doctrine of double effect.

Widespread norms of morality (like the prohibition of murder) can readily be understood in terms of the concept of the culpability of self. Inflicting unwarranted harm on others invites retaliation and that is detrimental to one's own survival chances. Moral values seem innate because, in our evolutionary past, they protected self from its own detrimental excesses. Norms of moral behavior do vary across societies, however, and we have argued that the social aspect of self—the existence of which is acknowledged and documented by social psychologists—responds to the sanctions imposed by society for

[51] In a nation-wide poll conducted by Gallup in the Unites States in May 2007, only 16% of the 1,003 adults whose opinion was polled said suicide is morally acceptable, whereas 49% felt that doctor-assisted suicide is morally acceptable. (http://www.pollingreport.com/values.htm)

violating its norms. These norms can be different, we have seen, depending on the extent to which the ecological niches occupied by different societies vary.

It has recently been argued that the moral codes espoused by many religions may be explained on an evolutionary basis [e.g. Hauser (2006, Ch. 7)]. Hauser points out that the behavior of people in moral dilemmas is independent of their religious beliefs. Atheists, agnostics, and those with religious faith all deliver the same judgments in the moral dilemmas he considers. Among religious people, it did not matter whether they were Catholics, Protestants, Jews, Sikhs, or Muslims—their verdicts tended to be the same. Hauser concludes (p. 421): "These observations suggest that the system that unconsciously generates moral judgments is immune to religious doctrine." This finding is compelling, and it comports well with our claim that it is the self-awareness of culpability that is responsible for our judgments in moral dilemmas. Being products of evolution the self and its awareness of culpability are common to all individuals (barring pathological cases), irrespective of their religious persuasions.

It is easy to see that the logic of our argument could easily be adapted to explain the sort of morality that is embodied in the Golden Rule, which says that we should treat others as we would have them treat us—a rule that has been proposed in many cultures.[52] This, however, is not the same thing as suggesting that *all* moral codes are consistent with natural selection or that the Universal Moral Grammar with a suitable setting of its parameters can replicate moral codes that have a religious basis. For example, the Christian ideal of treating one's neighbor as oneself is not likely to obtain in *any* society or culture as a product of evolution. For it is highly unlikely that natural selection in humans would lead to a complete obliteration of the self/other distinction. An ideal that strives to have us treat our neighbors as ourselves has to be based on a principle other than natural selection. Pervasive moral codes that we do observe may be seen as very modest steps taken in this direction by natural selection—to the extent that they were expedient for survival.

[52] Hauser's argument for this is based largely on reciprocity in long-term relationships. Our argument does not require long-term relationships.