Morality as an Emergent Property of Human Interaction

Henry Hinnefeld
Department of Physics
University of Illinois Urbana-Champaign

December 19, 2012

Abstract

Altruism directed towards non-related individuals has long presented an evolutionary puzzle: why would the intensely competitive process of natural selection favor individuals who helped their genetic rivals, occasionally disadvantaging themselves in the process? For many years, the prevailing answer in the scientific community was that altruism somehow directly benefited the altruist; that apparent displays of selflessness were in fact subtly self-interested. This belief stemmed from a reductionist approach to evolutionary biology, epitomized by Richard Dawkins' 1976 book "The Selfish Gene," which posited that all evolutionary analyses should be conducted at the level of the gene. Recently, however, a view based on a different level of description has been gaining acceptance. First suggested by Darwin himself in 1871, the idea of 'group selection', wherein natural selection happens at the level of groups, as well as individuals and genes, has recently found support in a variety of experiments. One of the most interesting implications of these experiments is that the morality underlying cooperation and non-kin altruism may be a result of group-level natural selection. When applied to questions of morality, these 'Multi-Level Selection Theories' imply that morality arises from interactions between humans within a group; that is, morality is emergent in collective human behavior. In this paper I will present the experimental evidence for this conclusion, as well as a historical summary of morality theories, from the reductionist theories prevalent in the past to current emergent models.

1 Introduction

Where does human morality come from? This question, for all its apparent simplicity, does not have a simple answer, as evidenced by the long and complicated history of thinking on the subject. Contemporary efforts to scientifically investigate morality are broadly focused in two fields, social psychology and evolutionary biology. Each of these disciplines addresses a different aspect of human morality; social psychology attempts to answer the question 'How does an individual acquire morals?', while evolutionary biologists ask 'How did humans acquire morals?' Tracing the interweaving strains of thought from each of these disciplines through their historical evolution reveals a surprising conclusion: morality is an emergent phenomenon in collective human behavior.

More precisely, morality is *very likely* to be an emergent phenomenon in collective human behavior. In both evolutionary biology and social psychology the relevant theories are the subjects of ongoing research; paraphrasing David S. Wilson [12], to examine theories of human morality is to see science in motion. Nonetheless, mounting evidence supports the idea that human morality stems from evolutionary pressure applied at the level of groups; that it is an emergent property of human interactions.

This subject does not immediately lend itself to analysis in a physics course, however upon closer inspection several of the key concepts discussed this semester make an appearance. In evolutionary biology the continuing debate about the origins of human morality has at its core a debate about levels of description. Meanwhile, in social psychology shared morals can introduce a broken symmetry in the interactions of genetically unrelated individuals.

2 Historical Overview of Morality Theories

2.1 Social Psychology

In social psychology, questions of morality often focus on how individuals acquire morals. As with evolutionary biology, the consensus within social psychology on the issue has varied considerably since the inception of the discipline in the early twentieth century. Early theories favored behavioral explanations, in keeping with the dominant paradigm in the social sciences at the time. In the late 1950s the advent of 'cognitive' sciences (neuroscience, artificial intelligence, cognitive psychology, etc), which emphasized the inner workings of the mind, sparked a shift to more explicitly rational theories of morality. Both models, and their apparent failings, are discussed below.

Ethical Behaviorism

Ethical Behaviorism emerged in the early 1900s as an attempt to make studies of the mind scientifically rigorous. Given the limited resources available for experimentally investigating the inner workings of the mind at the time, B. F. Skinner, J. B. Watson, and other early proponents of Behaviorism argued that since only behavior (and not any corresponding inner mental states or processes) could be observed, psychology should concern itself only with behavioral questions [4].

The concept of operant conditioning, wherein actions taken by an individual are positively or negatively reinforced as a result of their consequences, was central to behaviorist theories. When applied to questions of morality, this emphasis on conditioning and learned behavior led to the conclusion that a person acquired morals through the reinforcing effects of moral conduct. Ethical Behaviorism made no claims as to the origins of morality in broader human society, but at the individual level it claimed that morals are learned purely through positive and negative reinforcement; that "virtue is its own reward."

Starting in the 1950s, Behaviorism fell out of favor in the psychological and broader scientific community. Several causes can be found for its decline; first, the ability to probe inner mental states with modern technology has made the central tenet of Behaviorism (that as only behavior can be observed, only behavior should be considered in psychological explanations) obsolete. Second, the deterministic character of the theory, where behavior is determined uniquely by a prior history of reinforcement and punishment, was increasingly viewed as implausible.

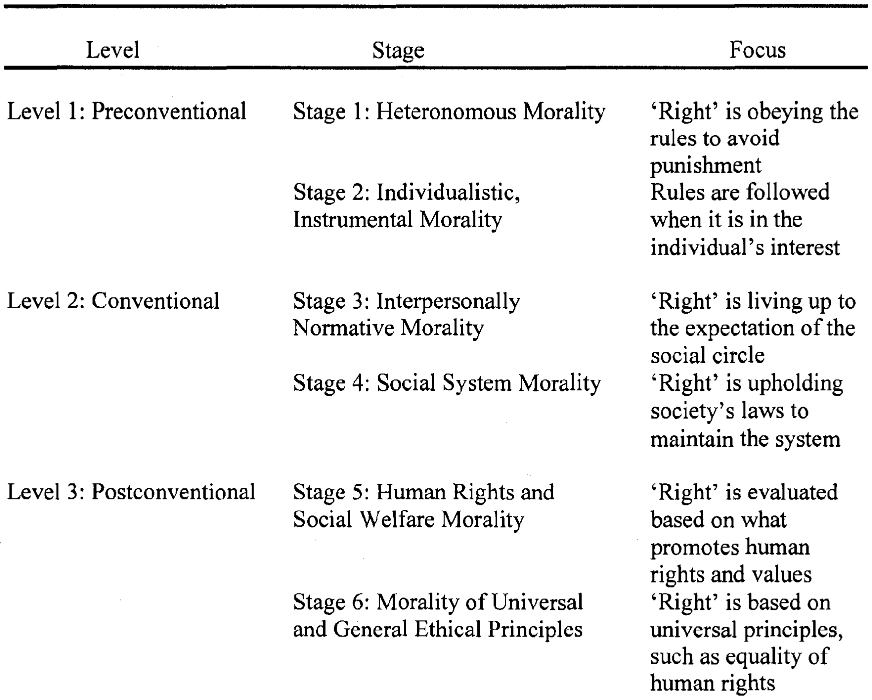
Kohlberg's Stages of Moral Development

In the late 1950s the consensus on morality in social psychology moved away from the behavioral theories of Watson and Skinner towards models that emphasized a rational, calculation-oriented approach to morality. Chief among these was the Stages of Moral Development theory developed by Lawrence Kohlberg starting in 1958. This theory claimed that behavior in moral situations was dependent on moral reasoning, and that the basis of this reasoning shifted through six stages during the course of a lifetime (as seen in Figure 1).

Kohlberg and his student Elliot Turiel developed the theory by describing moral dilemmas (such as the Heinz dilemma[1]) to subjects of various ages and recording their responses. By analyzing the reasoning behind the response, Kohlberg identified six basic modes of moral reasoning, as shown in Figure 1. Correlating the ages of the subjects with the level of reasoning used, Kohlberg defined a progression of increasingly complex bases for moral reasoning.

Recent experimental evidence calls into question Kohlberg's moral reasoning approach to morality. In several studies [5], Jonathan Haidt and collaborators have shown that moral judgements can exist in the absence of rational justifications. Through a series of questions involving morally repugnant, yet rationally unoffensive situations, Haidt et al. render subjects "morally dumbfounded." That is, subjects pronounce a moral judgement that they are unable to rationally justify through moral reasoning. A separate study involving hypnosis [6] likewise induced subjects to reach moral conclusions that they were unable to explicitly justify.

[1] In the form used by Kohlberg, "Heinz's wife was near death, and her only hope was a drug that had been discovered by a pharmacist who was selling it for an exorbitant price. The drug cost $20,000 to make, and the pharmacist was selling it for $200,000. Heinz could only raise $50,000 and insurance wouldn't make up the difference. He offered what he had to the pharmacist, and when his offer was rejected, Heinz said he would pay the rest later. Still the pharmacist refused. In desperation, Heinz considered stealing the drug. Would it be wrong for him to do that?"



Level	Stage	Focus
Level 1: Preconventional	Stage 1: Heteronomous Morality	'Right' is obeying the rules to avoid punishment
	Stage 2: Individualistic, Instrumental Morality	Rules are followed when it is in the individual's interest
Level 2: Conventional	Stage 3: Interpersonally Normative Morality	'Right' is living up to the expectation of the social circle
	Stage 4: Social System Morality	'Right' is upholding society's laws to maintain the system
Level 3: Postconventional	Stage 5: Human Rights and Social Welfare Morality	'Right' is evaluated based on what promotes human rights and values
	Stage 6: Morality of Universal and General Ethical Principles	'Right' is based on universal principles, such as equality of human rights

Figure 1: Lawrence Kohlberg's stages of moral development. Initially developed in the 1950s, the theory emphasizes conscious moral calculations, and the way the bases of such calculations vary during one's lifetime.

2.2 Evolutionary Biology

In evolutionary biology, morality is largely considered through the lense of altruism. Given its apparent evolutionary maladaptivity, altruism has attracted a considerable amount of scholarly interest, dating back to Darwin himself [2]. Group selection has played a varying part in theories of altruism; the history of thought on the issue brings to mind an underdamped oscillator: an initially overzealous and uncritical application of the theory during the first half of the 20th century led to its total dismisall in the 1960s. Only recently has a more nuanced view, which takes into account the reductionist theories described below, as well as more rigorous group selection arguments, emerged.

[2] In a famous passage from *The Descent of Man*, Darwin clearly laid out the arguments for group selection and its limitations: "It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage to one tribe over another. A tribe including many members, who from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection"

Kin Selection

The earliest attempts to explain altruism from a reductionist evolutionary persepective focused on how altruism might improve the fitness of an individual's genes by improving the fitness of other individuals having the same genes. Since direct relatives are guaranteed to share at least some portion of their genomes, inital evolutionary theories of altruism emphasized the relatedness of individuals engaged in altruistic behavior. Kin selection, as such theories came to be known, was given a mathematical formulation in 1964 by W. D. Hamilton [7], which gave rise to an equation describing the conditions under which altruistic behavior would evolve:

$$\text{Hamilton's Rule: } rB > C$$

where B is the benefit conferred on the recipient, C is the cost to the altruist, and r is a coefficient describing the relatedness of the two individuals.

Through the 1990s, kin selection found support in experimental evidence concerning the sex-determining mechanisms of eusocial (massively cooperating, and therefore altruistic) insects [8]. Specifically, until the 1990s nearly all known eusocial species were haplodiploid[3]; therefore sisters had an unusually high relatedness ($r = 3/4$). Given that the (altruistic) workers in such species are typically sisters, this was taken as evidence for kin selection as a mechanism for the evolution of altruistic behavior. However, starting in the 1990s many new non-haplodiploid eusocial species were discovered and the "Haplodiploid hypothesis" was largely abandoned [8]. While not fatal to kin selection theories, this setback, along with the difficulty of applying such theories to altruism beyond eusocial insects, eventually led to a broadening of the theory.

Reciprocal Altruism

Reciprocal altruism extends the core idea of kin selection, that altruistic acts are caused by specific genes and must somehow improve the fitness of those genes, to a broader class of situations. In kin selection, the benefit was applied to the the same genes in another individual, however in reciprocal altruism the benefit may also be applied to the inital altruist at a later time. The later benefit may be delivered by the initial recipient of altruistic behavior (so-called Direct reciprocity), or by a third party (Indirect reciprocity). In the latter case, altruistic behavior is motivated by reputation, where the likelihood of an individual receiving altruistic benefits is related to the perception of that individual's history of behaving altruistically.

Indirect reciprocity is particularly relevent to questions of human morality; it is a short step from "perception of altruistic behavior" to "perceived as a moral person." Research on the origin of langauage supports reciprocal explanations of morality; studies of primate social group size and neocortex ratio (the ratio of the size of the neocortex, which handles language, to the rest of the brain) show a strong correlation [2]. One possible explanation is that humans developed language in order to gossip [2]; that is, language was originally used to track reputations for indirect reciprocity.

[3] In haplodiploid species, males develop from unfertilized eggs and females develop from fertilized eggs.

A telling criticism levelled against both reciprocal altruism and kin selection theories (in [10] and elsewhere) is that they require genetic similarity to explain behavioral similiarity. While this is not a particularly unreaonable claim for the eusocial insects, humans and other organisms with more complex social interactions have correspondingly complex and variable behavior, even in situations with a high degree of genetic similarity. Alternatively, genetically diverse populations can exhibit similar behavior.

Inclusive Fitness Theory

The reductionist, gene-level theories described above were eventually consolidated into the larger framework of Inclusive Fitness Theory (IFT), one of the two currently competing theories in evolutionary biology. In IFT, altruism is still considered from a genetic level of description, however the fitness of a particular gene is expanded to include both a classical component (how many offspring share the gene) and a component that corresponds to the effects of the gene on copies of itself in other organisms. This second component is broadly interpreted to include non-descendent kin selection, reciprocal altruism, and other cooperative behavior effects.

Proponents of Inclusive Fitness Theory point to simulations, like the one shown in Figure 2, that show cooperation emerging in simulations where only individual-level selection effects are explicitly included. Detractors meanwhile emphasize the enormous complexity of human behavior, relative to the size of the genome, to argue that it is more appropriate to "study complexity at its own level" [3]. The ongoing debate is described in greater detail below.

3 Contemporary Morality Theories

3.1 Intuitive Primacy Morality Theories

Contemporary theories of morality in social psychology are moving away from the rational, moral calculation-oriented theories of the previous century towards theories stressing the intuitive, emotional[4] aspects of moral judgements [5], [6]. In particular, Jonathan Haidt and collaborators have developed a theory called Moral Foundations Theory (summarized in Figure 3) which emphasizes the intuitive primacy of moral judgements. According to the theory, moral judgements stem from emotion reactions, and the explicit reasons given for such judgements are *post hoc* rationalizations.

Haidt et al. propose a set of evolutionary bases for the emotions underpinning moral judgements. Research on morality in primates offers compelling evidence to support some of these claims. In particular, Frans de Waal's work with capuchin monkeys [1] illustrates that monkeys have a sense of fairness which prompts them to reject unequal pay for equal work. Other work by de Waal supports the premise of an evolutionary basis for empathy as well [9]. Finally, as mentioned above, studies into the origins of language [2] show a strong

[4] For an amusing visual example of 'emotional' reactions to moral slights see Frans de Waal's TED talk on capuchin monkeys and unequal pay: http://www.youtube.com/watch?v=g8mynrRd7Ak

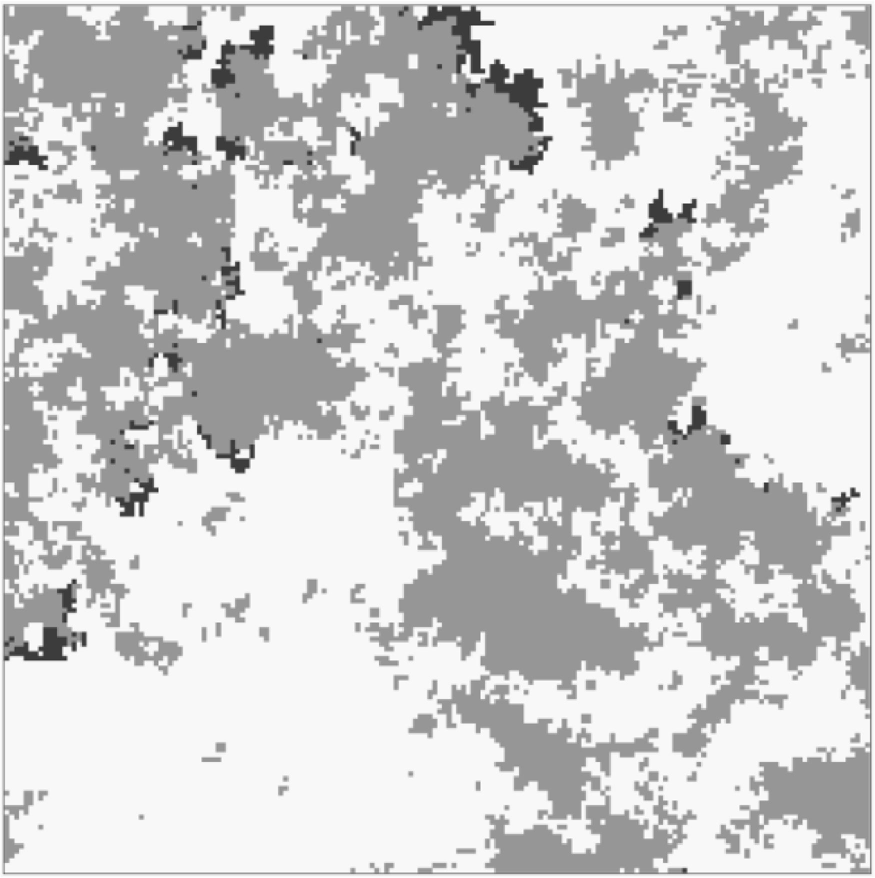


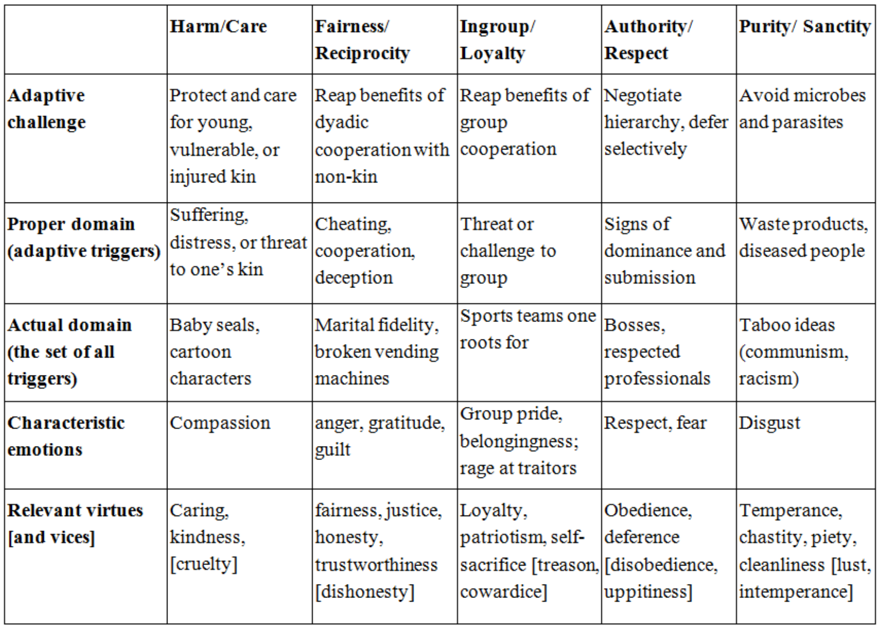
Figure 2: Distribution of cooperators (white), defectors (black), and loners (gray) on a 200 × 200 portion of a larger lattice, during a Public Goods Game (PGG) simulation. In PGGs, simulated organisms are placed at the vertices of a matrix and made to engage in pairwise Prisoner's Dilemma interactions with their neighbors. For certain conditions (e.g. the ability to opt out, the size of the reward for cooperating, the number of neighbors) cooperation emerges. Adapted from [11].

connection to the formation of social groups, which lends plausibility to the claim of an evolutionary basis for the Loyalty foundation.

3.2 Multi-Level Selection Theory

As described above, the idea of group selection was an academic anathema for the second half of the twentieth century. Recently, however, more nuanced group selection-based arguments are increasingly accepted within the context of Multi-Level Selection Theory (MLST). MLST presents an alternative to the gene-centric perspective of Inclusive Fitness Theory by allowing for natural selection at a variety of levels, including but not limited to the genetic. In contrast with IFT, MLST claims that selective pressure may be applied at the group level, and in such cases groups can be considered units of selection.

The emphasis within Multi-Level Selection Theory is on identifying the scale at which



	Harm/Care	Fairness/ Reciprocity	Ingroup/ Loyalty	Authority/ Respect	Purity/ Sanctity
Adaptive challenge	Protect and care for young, vulnerable, or injured kin	Reap benefits of dyadic cooperation with non-kin	Reap benefits of group cooperation	Negotiate hierarchy, defer selectively	Avoid microbes and parasites
Proper domain (adaptive triggers)	Suffering, distress, or threat to one's kin	Cheating, cooperation, deception	Threat or challenge to group	Signs of dominance and submission	Waste products, diseased people
Actual domain (the set of all triggers)	Baby seals, cartoon characters	Marital fidelity, broken vending machines	Sports teams one roots for	Bosses, respected professionals	Taboo ideas (communism, racism)
Characteristic emotions	Compassion	anger, gratitude, guilt	Group pride, belongingness; rage at traitors	Respect, fear	Disgust
Relevant virtues [and vices]	Caring, kindness, [cruelty]	fairness, justice, honesty, trustworthiness [dishonesty]	Loyalty, patriotism, self-sacrifice [treason, cowardice]	Obedience, deference [disobedience, uppitiness]	Temperance, chastity, piety, cleanliness [lust, intemperance]

Figure 3: Moral Foundations Theory, as initially developed by Jonathan Haidt and collaborators, seeks to explain moral judgements in terms of emotional responses. The theory claims that these emotional responses are based on five distinct triggers, each of which initially evolved to address a particular adaptive challenge. Adapted from [6].

the pressures of natural selection are being applied. In some cases, the appropriate scale is genetic, for example in the canonical example of beak shapes in finches. In other cases, however, proponents of MLST argue that one must consider selection operating at a larger scale in addition to individual-level effects. The question of human morality is a prime example of such a situation, as discussed below.

4 Relevance to Emergent States of Matter

4.1 MLST vs IFT as a Level of Description Debate

The debate within evolutionary biology between Multi-Level Selection Theory and Inclusive Fitness Theory is somewhat peculiar, from an outside perspective. Both sides claim that their preferred model is capable of expaining any phenomenon the other can, and to a large extent this is true. Why then is the issue so hotly contested? The vehemence of the argument stems from the desire on both sides to show that, not only does their model accurately explain evolutionary phenomena, it is in fact an accurate depiction of the evolutionary mechanisms behind them.

Here the two theories differ, and their difference lies in the level of description each favors. Inclusive Fitness Theory demands that all analyses be performed at the level of the gene, that since genes are the unit of inheritance they must be the unit of selection (though this premise is contested by those who consider culture to be an agent of inheritance, as in [10]). This is the reductionist position, advanced by Dawkins and decried by Gould [3]. Multi-Level Selection Theory, on the other hand, claims that selective pressure can be applied at many levels of description, including groups. That, as Darwin wrote in *The Descent of Man*,

> “a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage to one tribe over another . . . and this would be natural selection”

It is this fundamental disagreement about the appropriate level of description that separates the two theories.

Both theories claim the other contains logical fallacies, however one argument against Inclusive Fitness Theory is particularly relevant to the discussion of morals. The ‘Averaging Fallacy’ [12], claims that inclusive fitness calculations that attempt to maximize individual fitness often implicitly assume a group-level selective pressure. Consider the following example, adapted from [12]. Imagine two flocks of birds, one in which most birds give a warning call when a predator is spotted and one in which most do not. The act of giving a warning call reduces the chance of survival for the caller but increases the chance of survival for the rest of the flock. Choosing some values, say that in the first group callers have a 75 percent survival rate and noncallers have a 100 percent rate, while in the second group the rates are 25 and 50 percent. With a nine to one split between the common and uncommon behavior in each group, one can calcuate the average chance of survival; for callers it is 70 percent and for non-callers it is 55 percent. On the surface it appears that the calling behavior can be explained through selection at the individual level, however this calculation implicitly contains group-level effects in the differing survival rates between the two groups. The actual numbers are irrelevant, the point is that effects of group-level selection can become folded into apparently individual fitnesses by averaging over distinct groups.

That morality is a group-level phenomenon is undeniable; moral values like loyalty and reciprocity demand multiple parties by their very definitions. In considering morality from an evolutionary perspective one must therefore be careful to explicity account for group-level effects [12].

4.2 A Broken Symmetry Introduced by Shared Morals

Human morality introduces a broken symmetry in the interactions of unrelated individuals. Where two people might otherwise treat each other with equanimity, introduce either shared or competing moral values and a drastically different situation obtains. Examples abound, from the culture wars in American politics to the innumerable (admittedly more complex) conflicts where religion stands in for shared morals. Shared morals not only enable communal living in social groups, they also demarcate the boundaries of those groups.

In evolution, traits propagate because of their competitive advantage, competitive being the operative word. Shared morals as a mechanism for social cohesion are only evolutionary adaptive in the presence of natural selection between social groups [12]. In situations involving morality, the symmetry that one would expect in interactions between two random individuals is broken based on differences that did not exist *a priori*, but rather only once an individual adopted a particular set of moral values.

4.3 Morality as an Emergent Phenomenon

Having surveyed the history of thought on the issue, as well as the current scientific consensus, the emergent nature of human morality is apparent. Combining the perspectives of social psychology and evolutionary biology, the story that emerges is as follows: morality evolved by repurposing existing emotional triggers (as described in Figure 3) to promote cooperative and prosocial behaviors. The selective pressure of inter-group competition favored those groups whose intra-group interactions were modified by such intuitive moral reactions. Thus morality emerged from human interaction.

Human morality does not stem from an intrinsic moral faculty, the intuitive bases of moral judgements notwithstanding. It is interactions with other humans, and human culture, that shape raw emotional responses into moral codes [10]; therefore only by considering group-level effects can the evolutionary adaptivity of morality be accurately assessed. Until recently, this was not a popular position. However, with the increasing emphasis on emergent phenomena in the broader scientific community, the consensus is beginning to shift away from the reductionist theories of the past towards more complete theories that embrace the inherently collective nature of human morality.

References

[1] Brosnan, Sarah F., and Frans de Waal. “Monkeys reject unequal pay.” *Nature* 425.6955 (2003): 297-299.

[2] Dunbar, Robin. *Grooming, Gossip, and the Evolution of Language.* Cambridge, MA: Harvard University Press, 1993.

[3] Gould, Stephen J. “Humbled by the Genome’s Mysteries.” *New York Times* 19 Feb. 2001

[4] Graham, George. “Behaviorism.” *The Stanford Encyclopedia of Philosophy.* Ed. Edward N. Zalta. http://plato.stanford.edu/entries/behaviorism/

[5] Haidt, Jonathan. “The New Synthesis in Moral Psychology.” *Science* 316.5827 (2007): 998-1002.

[6] Haidt, Jonathan. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* New York, NY: Pantheon, 2012.

[7] Hamilton, William D. “The genetical evolution of social behaviour I.” *Journal of Theoretical Biology* 7.1 (1964): 1-16.

[8] Nowak, Martin A., Corina E. Tarnita, and Edward O. Wilson. "The evolution of eusociality." *Nature* 466.7310 (2010): 1057-1062.

[9] Preston, Stephanie D. and Frans de Waal. "Empathy: Its ultimate and proximate bases." *Behavioral and Brain Sciences* 25.1 (2002): 1-20.

[10] Richerson, Peter J., and Robert Boyd. *Not By Genes Alone: How Culture Transformed Human Evolution.* Chicago, IL: University of Chicago Press, 2005.

[11] Szabó, György, and Christoph Hauert. "Phase Transitions and Volunteering in Spatial Public Goods Games." *Physical Review Letters* 89.11 (2002): 118101.

[12] Wilson, David S. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society.* Chicago, IL: University Of Chicago Press, 2003.

[13] Wilson, David S., and Edward O. Wilson. "Rethinking the theoretical foundation of sociobiology." *Quarterly Review of Biology* 82.4 (2007): 327-348.