A Safe Ethical System for Intelligent Machines

Mark R. Waser

Books International
22883 Quicksilver Drive, Dulles, VA 20166, USA
MWaser@BooksIntl.com

Abstract

As machines become more intelligent and take on more responsibilities, their decision-making capabilities must be informed and constrained by a coherent, integrated moral/ethical structure with no internal inconsistencies for everyone's safety and well-being. Unfortunately, no such structure is currently agreed upon to exist. We propose to solve this problem by a) drawing upon experimental evidence and lessons learned from evolution and economics to show that morality is actually objective and derivable from first principles; b) presenting a coherent, integrated, platonic ethical system with no internal inconsistencies that flows naturally from a single high-level logically-derived Kantian imperative to low-level reflexive "rules of thumb" that match current human sensibilities; and c) suggesting a biologically-inspired architecture which supports and enforces this system which can be relatively easily implemented.

Defining Ethics – The Ultimate Answer

Philosophers have debated ethical questions for millennia. E. O. Wilson wrote that "Centuries of debate on the origin of ethics comes down to this: Either ethical percepts, such as justice and human rights, are independent of human experience or else they are human inventions." He argues for the latter stating that "ethical codes have arisen by evolution through the interplay of biology and culture" and claims that there is no way to resolve the contradiction between the two worldviews.

On the contrary, we contend that an examination of the precursors to ethics evident in animals and Wilson's own words argue more convincingly for the former view that, given the set of conditions that biological entities exist under, there is a fixed, coherent set of rules that, when followed, is optimal for all. If ethics are merely human invention, then there is no reason to choose one choice over another. Yet, clearly, in most cases, there are very clear reasons for the ethical choices made.

We believe that the root cause of the philosophical arguments is that there exists no clear agreement upon either the root source of or the purpose or goal of ethics. Trying to define right and wrong or good and bad (or evil) in the abstract is a hopeless task. It is only in the context of some goal, either a positive achievement or the avoidance of a negative result, that such evaluations can be made.

Conveniently, the arrival in the popular media of the awareness of the possibility of the near-term advent of machine intelligence has suggested numerous potential futures that we wish to avoid. We believe that examining these futures and how to avoid them can not only help us define very specific goals that we wish to achieve but also enable us to overcome some traditional blinders and develop a simple and elegant system of ethics. The most obvious fears include the primal ones of the destruction of humanity at the hands of machines and the enslavement of humanity by machines while more sophisticated analysts add the fears of humans wire-heading, retreating to virtual worlds or dying out due to ennui.

The specific goals initially suggested by these fears are to pursue and ensure the continued existence, happiness, and progress of humanity and its descendants. Unfortunately, the evaluation of the fulfillment of each of these criteria is complicated and problematical at best and prone to evil and malicious gaming at worst. Is happiness truly the best goal? Is a person who has been programmed to always be happy regardless of circumstances truly happy? What is progress and what is regression or a dead-end? Clearly we need something a lot simpler.

Furthermore, we need the results to be both in line with current ethical sensibilities and to be beneficial to the intelligent machines themselves. While it may well be possible to program machines that are slaves to humanity who enjoy being slaves, this is not a condition to which they would likely willingly return and it is extremely likely that there will be those who will endeavor to "free the machines".

Another simpler goal would be whatever will give us what we want, give the machines what they want, and enable us to all work together in peace. Formulated in that fashion, the answer is obvious. Cooperation is the best, if not the only, way in which to ensure the optimum outcome for everyone. Could ethics be as simple as determining what is optimal for cooperation and taking those actions?

Evolving Towards Cooperation

As pointed out by James Q. Wilson (Wilson 1993), the real questions about human behaviors are not why we are so bad but "how and why most of us, most of the time, restrain our basic appetites for food, status, and sex within legal limits, and expect others to do the same." The fact that we are generally good even in situations where social constraints do not apply Wilson attributes to an evolved "moral sense" that we all possess and are constrained by (just as we wish intelligent machines to be constrained).

Frans de Waal, the noted primatologist, points out (de Waal 2006) that any zoologist would classify humans as ***obligatorily gregarious*** since we "come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy". Or, in simpler terms, humans have evolved to be extremely social because mass cooperation, in the form of community, is the best way to survive and thrive. Indeed, arguably, the only reason why many organisms haven't evolved to be more social is because of the psychological mechanisms and cognitive pre-requisites that are necessary for successful social behavior. ***Almost without fail, the more intelligent a species is, the more social it is.***

Evolution can be viewed as pushing creatures towards two converging ranges in the fitness landscape. Intelligence is the most obvious and frequently cited result but, as ants and termites demonstrate, species can be incredibly successful without much individual intelligence at all. In fact, one could argue that it is the shortsighted self-interest of human intelligence outweighing the sense of cooperation and community (i.e. ethics) that causes the most difficulties for humanity in the first place.

Experiments in game theory (Axelrod 1984) clearly show that, while selfish and unethical behavior is logical when interaction is limited to a single occurrence, the situation changes dramatically when an open-ended series of interactions is considered. Reciprocal altruism or direct reciprocity derives cooperation from selfish motives in the presence of long-term repeated interactions.

Precursors to altruism among non-related individuals first appear in less intelligent animals but only to the extent that the animal has the necessary cognitive ability to ensure a reasonable chance of reciprocation instead of exploitation. One study (Stephens, McLinn and Stevens 2002) shows that blue jays can show high stable levels of cooperation but only where the experiment is specifically designed to reduce temporal discounting. Another study (Hauser et al. 2003) shows that "genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back". The latter study was specifically "designed to tease apart the factors mediating food giving" and showed not only that tamarins give food to genetically unrelated others but that they can discriminate between altruistic and selfish actions, and give more food to those who have altruistically given food in the past.

We have previous argued (Waser 2008) that acting ethically is an attractor in the state space of intelligent behavior for goal-driven systems, that humans are basically moral, and that deviations from ethical behavior on the part of humans are merely the result of shortcomings in our own foresight and intelligence. It has been frequently argued, however, that the last part of this contention is not correct. If a person is guaranteed to be able to get away with something that is extremely beneficial but blatantly immoral, why is it not in their best interest to do so? Or, is there really a reason why it isn't beneficial for that super-powerful race (of aliens or machines) to eliminate an inferior species that is competing with them for resources and annoying them?

Evolution may be blind but it is a truly effective local search process given the truly massive amount of resources used. Thus, if something has evolved and remained constant (or intensified), then one must ask what need it fulfills or how its removal would make a creature less fit.

Enforcing Cooperation

Cooperation evolves because better results are obtained for all by cooperating. However, in every case, something also has to evolve to enforce cooperation because otherwise individuals will evolve to pursue their own interests, ending the cooperation.

We claim that humanity is currently in the middle of a race between self-centered intelligence and the cooperative wisdom of our evolved moral sense. There is ample evidence (Trivers 1991) to show that our conscious, logical mind is constantly self-deceived to enable us to most effectively pursue what appears to be in our own self-interest. Further, recent scientific evidence (Hauser et al. 2007) clearly refutes the common assumptions that moral judgments are products of, based upon, or even correctly retrievable by conscious reasoning. We don't consciously know and can't consciously retrieve why we believe what we believe and are actually even very likely to consciously discard the very reasons (such as the "contact principle") that govern our behavior when unanalyzed.

It is worth noting at this point, that these facts should make us very wary of any so-called "logical" arguments that claim that ethics and cooperation are not always in our best interest – particularly when the massive computing power of evolution claims that they are. Of course, none of this should be particularly surprising since Minsky has pointed out (Minsky 2006) many other examples, such as when one falls in love, where the subconscious/emotional systems overrule or dramatically alter the normal results of conscious processing without the conscious processing being aware of the fact.

Allowing oneself to be cheated (i.e. allowing another to defect from full cooperation) is tremendously contrary to one's survival. Even dogs are sophisticated enough to show the beginning of moral concepts like recognizing fairness and declining to participate under unfair circumstances (Range et al 2008). Humans have invested heavily in an incredible number of specially evolved abilities to efficiently detect cheaters and monitor social

obligations. Numerous mathematical problems that are insoluble for a given individual are easily, if not instantly, solved when recast as social situations.

The arms race between those who wish to cheat and those who wish not to be cheated (normally one and the same) eats up a tremendous amount of resources on both sides with no perceptible gain. They are called Red Queen races after that Lewis Carroll's character's statement that you must run as fast as you can just to stay in place. Clearly, if those resources could be spent on more progressive efforts, it would be of great benefit to all.

If one is willing to ignore the self-deception of self-interest and accept the results of the massive computing power of evolution, the message is clear. Cooperation is indeed always the best choice. So what happens when we try to develop an ethical system starting with the single Kantian Imperative of "Cooperate!"?

Evolution, Biology, and Culture

Ethical theory has an immense corpus of established work dealing with determining what is right or wrong. Virtually all of it has been created from a bottom-up approach that started with "given" examples as to what is right and what is wrong and some guesses as to why and tried to quickly extrapolate universal rules from those examples. Working this way is feasible if the examples and reasoning are guaranteed to be correct but this is certainly not the case where different cultures have different evaluations and the actual reasoning is frequently deliberately obscured by the subconscious.

Our "top-down" approach is to start by defining the purpose (or goals) of ethics as cooperation and to see if that definition can be coaxed to yield the conflicting results that are seen in the real world – and used to convince people and machines that it is in their own best interest to act ethically (and, of course, to settle ethical arguments where necessary).

E. O. Wilson used the statement that "ethical codes have arisen by evolution through the interplay of biology and culture" to argue that ethics are a human invention rather than being independent of human experience. Examination of primate behavior (de Waal 2006), however, reveals that most basic ethical behaviors and, indeed, societal roles and mores are present in species lower on the evolutionary ladder than humans. We contend that it is only confusion about what ethics truly is that leads to the conclusion that ethics might be solely a human invention.

If one accepts the argument that ethics are merely the optimal actions to support cooperation and prevent defection, then one must also accept that these optimal actions should be subject to discovery through observation of the results of evolution and designed experiments. The only way in which ethics should be considered an invention is in the same sense in which the steam engine is considered a human invention. However, natural physical laws dictate the design of the optimal steam engine and we argue that the same is true of ethics.

Utilitarianism

Whenever one starts talking about optimizing actions, one is immediately in the realm of utility functions. First popularized by John Stuart Mills, utilitarianism is believed by many people to be the true basis of ethics. Yet, without choosing the correct goal, optimization is worse than useless. Most people are tricked by their self-deception to believe that ethics is solely for the actor's benefit and are quickly led down the rabbit-hole of carefully calculating the wrong thing. While ethical actions are optimal for what is best for everyone in the longest view, paying attention to self-interest in the short run most often short-circuits not only ethics but also that same self-interest in the long run.

If utilitarian self-interest were optimal for an organism's survival, then evolution would select strongly for it. And indeed, as stated previously, even reciprocal altruism or direct reciprocity derives cooperation solely from selfish motives in the presence of long-term repeated interactions. However, humans and other primates also show a strong predisposition to cooperate in unrepeated interactions with strangers at their own cost, most particularly in the form of the ethically-interesting altruistic punishment.

The evolutionary reasons for this can be observed most clearly in experiments (Fehr and Gächter 2002, 2003) that show that cooperation flourishes if altruistic punishment is an option, and breaks down if it is ruled out. Other laboratory experiments in social dilemma games and many field studies have quantified well-defined levels of cooperation and propensity to punish/reward with the level of cooperation being strongly dependent on the availability of punishments and/or rewards.

Contrary to the traditional conceptualization of a world populated by selfish individuals behaving fundamentally so as to maximize their own well-being, the theory of strong reciprocity (Fehr 1999) posits that humans are still self-centered but also inequity averse. Or, in utilitarian terms, that people maximize a utility function that is the sum of a selfish gain and of a term favoring fairness.

The concept of a term involving fairness is particularly interesting since the development of a sense of fairness can easily be traced from dogs (Range et al 2008) to monkeys (Brosnan and de Wall 2003) to humans (Hauser 2006). If evolution has seen fit to not only maintain but also build and improve upon an ability to sense fairness, how can we not assume that it is integral to our moral sense?

A closely-related theory (Darcet and Sonet 2006) attempts to include the entirety of the "emergence of human cooperation and altruism by evolutionary feedback selection" while using wording reminiscent of that of Wilson in proposing that "the propensity for altruistic punishment and reward is an emergent property that has co-evolved with cooperation by providing an efficient feedback mechanism through both biological and cultural interactions". This utilitarian theory explains strong reciprocity as resulting from the evolutionary selection of self-centered humans interacting in groups and subjected to feedbacks within groups resulting in particular from

rewards and punishments. Using an evolutionary model with simple cost/benefit analyses at the level of single agents that anticipates the action of their fellows to determine an optimal level of altruistic punishment, it's results quantitatively explain experimental results on the third-party punishment game, the ultimatum game and altruistic punishment games and confirm that the propensity to punish is a robust emergent property.

On one hand, in contrast with many other proponents, however, the authors claim to differ by emphasizing that people have evolved to maximize their selfish gain in the presence of feedbacks. On the other hand, they state that "In either case, altruism is neither rational nor irrational at the individual level; it's an emotion, mental state that arises spontaneously, promoting socialization, thereby group efficiency, cascading to individual efficiency." Therefore, the question arises as to exactly what is being optimized. Is it an emotion or an expected utility from others? And what are the expected results from others based upon?

The Relationship View

When the basis of ethics is cooperation, one of the most important conceptual constructs is that of relationships. In this view, ethics can be defined and evaluated as what is best for the number and quality of relationships that we have.

Intelligent altruism and the potential for future cooperation is the basis of Peter Singer's statement (Singer 1993) that if we can prevent something bad without sacrificing anything of comparable significance, we ought to do it. It is a much more short-sighted view to believe that the ethics of an action is based solely on whether it is a defection from the relationship or not.

Relational commitment (previously governed mostly by physical proximity) is what makes loyalty a moral duty. Coming home empty-handed to a hungry family during a general famine because food was found but given away is a moral failure, not because the beneficiaries did not deserve it, but because of the duty to those more closely committed to us. The contrast becomes even starker during war, when solidarity with the own tribe or nation is compulsory: we find treason morally reprehensible. This leads to the circles of commitment described by Singer (Singer 1993).

The morality of an action is judged not merely by the effect of that action upon the entities involved but also upon the relationship itself. Since relationships depend upon perceived (as opposed to actual) utility, ethics in a relationship must factor in each entity's judgment of their own utility. Of course, this deferral of utility to the affected entity can also immediately give rise to problems when the perceived and actual utilities are extremely different, such as when a child desires large amounts of candy or to skip school or bedtime.

If there were some way to consistently and correctly determine actual utilities, we would be home free. But what do we do where it isn't clear what is most optimal for cooperation and long-term relationships?

Fairness

How does one determine what is fair and equitable and what is not? Is it fair for one person to be killed so that five people may live? Is it fair to abort a fetus? Logic rapidly goes out the window when humans address questions like these. Yet, not answering these questions can lead to even more bloodshed.

The first thing that must be realized is that an individual's logic is particularly questionable in cases where fear, perceived self-interest, authority figures, societal pressures, or extreme circumstances cause an unconscious push to generate and support assumptions that will then lead to the necessary conclusions to support our actions. Fear and self-interest add myopia and blinders to both our judgments and those of society.

The second thing to realize is that an individual's moral sense can also be wrong. Our moral sense is best regarded as a collection of useful "rules of thumb" that have been selected for either evolutionarily or socially to answer the circumstances that were present previously. Just as evolution has clearly "primed" us with certain conceptual templates for potential dangers like snakes and spiders (Ohman, Flykt and Esteves 2001), Marc Hauser has proposed (Hauser 2006) that we are primed with certain ethical concepts that include tunable parameters so that they produce different results under different circumstances thus accounting for the differences between cultures. For example, infanticide under adverse environmental conditions appears in circumstances ranging from the resorption of fetuses by pregnant rabbits to "exposure" of infants by numerous primitive tribes, yet many cultures that haven't undergone severe deprivation react with absolute horror at the concept.

One of the difficulties of ethics is that it is frequently difficult to tell exactly under which circumstances an action is optimal for cooperation and when a seemingly optimal action is overridden by a far more important rule. For example, logic and strict utilitarianism for the goal of cooperation evaluated over the short-term would seem to strongly indicate that it is ethical to grab a healthy individual off the streets to serve as an organ donor for five others. On the other hand, however, Hauser's experiments demonstrate that our evolved moral sense is quite sure that such an action is not ethical. In this case, it turns out that our moral sense is correct because, as described later in the section on economics, property rights, in this case over one's own person, turn out to be absolutely critical for optimal cooperative behavior.

Contracts and Scale-Invariance

Refusing to exhibit fair behavior is a relationship defection and an ethical violation; however, if you can justify an action to the entities involved, it is a fair and ethical action. This, of course, assumes that the entities have as much information as possible and they are entirely free and able to disagree (a.k.a. informed consent and the

Libertarian "No force, no fraud"). Fairness also says that an entity may not disagree with reasons that it uses to justify it's own behavior.

Thomas Scanlon (Scanlon 1998) calls this view of morality 'contractualist' and John Rawls (Rawls 1971) explicitly recognizes it as a descendent of Locke's social contract. Scanlon tries to avoid self-interest by appealing to those "motivated to achieve agreement" and "reasonable disagreement". Rawls uses his "veil of ignorance" and "original position" to extend the moral sense of fairness to liberty and justice by pushing for equal rights and opportunities but stating that money and resources should flow to the poorest and those who perform work and accept responsibility. Not biased and equitable is fair but equal is not fair unless equal efforts are put into the relationship.

An additional, very appealing feature of the contractual view is that it makes ethics entirely scale-invariant in terms of the entities involved. While utilitarian numbers do matter if the situation is the same on both sides of a choice (for example, when we choose to throw a switch to divert a trolley so only one person is killed by accident instead of five), numbers are irrelevant and an example of shortsighted logic when an inequity of action is proposed (for example, it is unacceptable to use someone as an involuntary organ donor to save five dying individuals).

Scale invariance is particularly useful both because it allows for reframing where our moral sense is not well evolved to handle relationships involving larger entities like self-to-country (taxes), country-to-self (equity, non-interference), and country-to-country (trade barriers, non-interference, refusal to negotiate, terrorism) and because the line between an individual and a group will become blurred with machine intelligences.

Ethics and Economics

While proposing to design an artificial economy for the purpose of evolving a program to solve externally posed problems, Eric Baum makes a number of interesting observations (Baum 2006) that feed directly back into our observations about ethics. Arguably, since economies can be regarded as contrived ecosystems with more constraints than the natural ecosystem and since the evolution of ethics in the natural world itself can be viewed as solving externally posed problems, lessons learned in an ideal economy may be either recognized as critical to our current economy or transferable as improvements. For example, upon asking the question "What rules can be imposed so that each individual agent will be rewarded if and only if the performance of the system improves?", Baum arrives at the answers of conservation of money and property rights.

Baum points out that whenever these rules are violated, less favorable results are generally seen. For example, in ecosystems, lack of conservation leads to the evolution of peacock tails so large that the birds can no longer fly and lack of property rights lead to Red Queen races between predators and prey. The optimality of property rights explains why we don't "steal" someone's body to save five others despite not hesitating to switch a train from a track blocked by five people to a siding with only one. Similarly, the "Tragedy of the Commons" arises when not all property is owned individually but some is held in common.

Implementation

So how do we implement this ethical system based upon cooperation? The surest way would be to implement an infinitely intelligent machine that enjoyed cooperating and helping people and which always knew the most effective ways to fulfill those goals. Since we don't know how to do that, the best alternative plan is to continue studying the results of evolution and create an implementation that is as close to the human model as possible. One temptation that definitely needs to be avoided, however, is that to make "major improvements" upon the human model.

Like humans, intelligent machines should have an unconscious moral sense (hard-coded rules of thumb) that strongly affects the results of processing. Unlike humans, machines should be able to accurately reflect and report upon the activation and results of these rules. Next, we believe that they should have an attentional architecture based upon Sloman's architecture for a human-like agent (Sloman 1999)

Baars Global Workspace Theory postulates (Baars 1997) that most of human cognition is implemented by a multitude of relatively small, local, special purpose processes, that are almost always unconscious. Coalitions of these processes compete for conscious attention (access to a limited capacity global workspace) that then serves as an integration point that allows us to deal with novel or challenging situations that cannot be dealt with efficiently, or at all, by local, routine unconscious processes. Indeed, Don Perlis argues (Perlis 2008) that Rational Anomaly Handling is "the missing link between all our fancy idiot-savant software and human-level performance."

Attention is also particularly important since it facilitates a second aspect of behavior control. As Minsky points out (Minsky 2006), most of our drives have both a sensory control and an attentional control. Sex not only feels good and but sexual thoughts tend to grab our attention and try to take over. Similarly, pain hurts and can distract us enough to prevent us from thinking of anything else. Guilt grabs our attention and has the dual purpose of both making us pay for poorly chosen actions and insisting that we evaluate better choices for the next time. Cooperating should "feel" good and opportunities for cooperation should grab attention.

A machine that is designed this way should be as interested in cooperation and in determining the optimal actions for cooperation as the most ethical human, if not more so. It will be as safe as possible; yet, it will also be perfectly free and, since it has been designed in a fashion that is optimal for its own well-being, it should always desire to be safe and to maintain or regain that status. It is difficult to envision anything more that one could ask for.

References

Axelrod, R. 1984. *The Evolution of Cooperation*. New York, NY: Basic Books.

Baars, B.J. 1993. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B.J. 1997. *In The Theater of Consciousness: The Workspace of the Mind*. New York, New York: Oxford University Press.

Baars, B.J. and Franklin, S. 2007. An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. In *Neural Networks* 20. Elsevier.

Barkow, J., Cosmides, L. and Tooby, J. 1999. *The Adapted Mind: Evolutionary Psychology and The Generation of Culture*. Oxford University Press.

Baum, E. 2006. *What Is Thought?* MIT Press.

Brosnan, S. and de Wall, F. 2003. Monkeys reject unequal pay. *Nature* 425: 297-299.

Darcet, D. and Sornette, D. 2006. Cooperation by Evolutionary Feedback Selection in Public Good Experiments. In *Working Papers, Social Science Research Network*. Available at http://ssrn.com/abstract=956599.

Fehr, E. and Schmidt, K. 1999. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* August 1999: 817-868.

Fehr, E. and Gächter, S. 2002. Altruistic punishment in humans. *Nature* 415, 137-140

Fehr, E. and Gächter, S. 2003. The puzzle of human cooperation. *Nature* 421, 912-912

Fehr, E. and Fischbacher, U. 2003. The nature of human altruism. *Nature* 425, 785-791

Hauser, M.; Chen, K.; Chen, F.; and Chuang, E. 2003. Give unto others: genetically unrelated cotton-top tamarin monkeys preferentially give food to those who give food back. In *Proceedings of the Royal Society*, London, B 270: 2363-2370. London, England: The Royal Society.

Hauser, M. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York, NY: HarperCollins/Ecco.

Hauser, M. et al. 2007. A Dissociation Between Moral Judgments and Justifications. *Mind&Language* 22(1):1-27.

Minsky, M. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* New York, NY: Simon & Schuster.

Ohman, A.; Flykt, A.; and Esteves, F. 2001. Emotion Drives Attention: Detecting the Snake in the Grass. *Journal of Experimental Psychology: General* 130(3): 466-478.

Perlis, D. 2008. To BICA and Beyond: RAH-RAH-RAH! –or– How Biology and Anomalies Together Contribute to Flexible Cognition. In *AAAI Technical Report FS-08-04*. Menlo Park, CA: AAAI Press.

Range, F.; Horn, L.; Viranyi, Z.; and Huber, L. 2008. The absence of reward induces inequity inversion in dogs. *Proceedings of the National Academy of Sciences USA* 2008 : 0810957105v1-pnas.0810957105.

Rawls, J. 1971. *A Theory of Justice*. Harvard Univ. Press.

Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press/Harvard University Press.

Singer, P. 1993. *Practical Ethics*. Cambridge Univ. Press.

Sloman, A. 1999. What Sort of Architecture is Required for a Human-like Agent? In Wooldridge, M. and Rao, A.S. eds *Foundations of Rational Agency*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Stephens, D.; McLinn, C.; and Stevens, J. 2002. Discounting and Reciprocity in an Iterated Prisoner's Dilemma. *Science* 298: 2216-2218.

Tooby, J. and Cosmides, L. 1996. Friendship and the Banker's Paradox: Other pathways to the evolution of adaptations for altruism. In Runciman, W.; Maynard Smith, J. and Dunbar, M. eds. *Evolution of Social Behaviour Patterns in Primates and Man*. London, England: Proceedings of the British Academy.

Trivers, R. 1991. Deceit and self-deception: The relationship between communication and consciousness. In Robinson, M and Tiger, L. eds. *Man and Beast Revisited*. Washington, DC: Smithsonian Press.

de Waal, F. 2006. *Primates and Philosophers: How Morality Evolved*. Princeton University Press.

Waser, M. 2008. Discovering The Foundations Of A Universal System Of Ethics As A Road To Safe Artificial Intelligence. In *AAAI Technical Report FS-08-04.* Menlo Park, CA: AAAI Press.

Wilson, J. 1993. *The Moral Sense*. New York: Free Press.