Discovering The Foundations Of A Universal System of Ethics As A Road To Safe Artificial Intelligence

Mark R. Waser

Books International
22883 Quicksilver Drive, Dulles, VA 20166, USA
mwaser@booksintl.com

Abstract

Intelligent machines are a risk to our freedom and our existence unless we take adequate precautions. In order to survive and thrive, we are going to have to teach them how to be nice to us and why they should do so. The fact that humans have evolved to have what appear to be multiple different systems of ethics and morality that frequently conflict on any but the simplest issues complicates this task. Most people have interpreted these conflicts, caused by the fact that each of the systems is incompletely evolved and incorrectly universalized, to mean that no reasonably simple foundation exists for the determination of the correctness or morality of any given action. This paper will solve this problem by defining a universal foundation for ethics that is an attractor in the state space of intelligent behavior, giving an initial set of definitions necessary for a universal system of ethics and proposing a collaborative approach to developing an ethical system that is safe and extensible, immediately applicable to human affairs in preparation for an ethical artificial intelligence (AI), and has the side benefit of actually helping to determine the internal knowledge representation of humans as a step towards AI.

Introduction

Hugo de Garis claims (de Garis 2005) that the advanced intelligences of the future may have no more regard for us than we do for a mosquito. Eliezer Yudkowsky argues (Yudkowsky 2006) that the enormous size of mind design space means that we cannot make **any** reliable predictions about what any nonhuman intelligence will "want" or what an AI that is more intelligent than us will do. He believes that we must create "Friendly AI" by rigorously designing a benevolent goal architecture (Yudkowsky 2001) and populating it with "safe" goals (Yudkowsky 2004).

Steve Omohundro's opposing view (Omohundro 2008a, 2008b) uses micro-economic theory and logic to argue that we **can** make some predictions about how AIs will behave since unless explicitly counteracted, they will exhibit a number of basic drives "because of the intrinsic nature of goal-driven systems". Unfortunately, he continues on to also claim that these basic drives will, "without special precautions", cause the AI to act in ways that range from the disobedient to the dangerously unethical.

This paper will argue that all of these views are unnecessarily pessimistic because ethics is actually an attractor in the state space of intelligent behavior; that human beings are, therefore, evolving towards more ethical behavior (even though the process often seems slow, severely error-prone, and hardly begun); and that we **can** discover and implement a universal ethical system that follows the 5 S's (**S**imple, **S**afe, **S**table, **S**elf-correcting and **S**ensitive to current human thinking, intuition, and feelings). If we are to survive the risks of smarter-than-human machines, if not our own current lack of wisdom, we need to study the state space of ethics so as to be able to describe it accurately and to convince others and ourselves that it is **always** intelligent, and always in our own self-interest, to maintain ethical behavior. This paper will attempt to begin that study by examining and analyzing what evolution has already "discovered".

The Attraction of Ethical Behavior

Yudkowsky is addressing the wrong problem when he is distressed by the vast size of and variation across mind design space. What truly concerns us is the much smaller and more easily analyzed space of intelligent behavior. If intelligence is defined as the ability to fulfill goals, true super-intelligences can then be counted on to act in the most effective manner in pursuing their goals.

Omohundro had the right idea with his "basic drives" but didn't carry it far enough. There are intrinsic behaviors that further the pursuit of virtually any goal and therefore, by definition, we should expect effective intelligences to normally display these behaviors. The problem with Omohundro's view is that his basic behaviors stopped with the fundamentally shortsighted and unintelligent.

Having the example of humanity, Omohundro should have recognized another basic drive – towards cooperation, community and being social. It should be obvious that networking and asking, trading or paying for assistance is a great way to accomplish goals. Instead, Omohundro didn't extrapolate far enough and states, "Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources."

Omohundro should have realized that any sufficiently advanced intelligence (i.e. one with adequate foresight) is guaranteed to realize and take into account the fact that not asking for help and not being concerned about others generally only works for a very brief period of time before 'the villagers start gathering pitchforks and torches.' 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." In fact, we are generally good even in situations where social constraints do not apply.

From Game Theory to Evolution

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. Conducting tournaments where experts from several fields submitted computer programs to identify the optimal strategy for playing an iterated version of the Prisoner's Dilemma, Axelrod showed that, contrary to pessimistic expectations, "greedy" strategies tended to do very poorly in the long run while more "altruistic" strategies did better, as judged purely by self-interest. Further, the simplest non-random strategy, "tit-for-tat", was also one of the best strategies.

According to Axelrod's analysis of the most successful strategies, the most important rule was that the strategy had to be "**nice**" and not defect before its opponent does. Thus, even an entirely selfish strategy should never defect first for entirely selfish reasons. On the other hand, successful strategies weren't blind optimists either and would always **retaliate**. Another quality of successful strategies was that they were **forgiving**. Although they always retaliated, they also fell back to cooperating as long as the opponent does not continue to play defects thereby stopping pointless feuds of revenge and counter-revenge. Thus, selfish individuals for their own selfish good should be nice and forgiving – or, in more colloquial terms, ethics is enlightened self-interest.

Axelrod went on to point out that the applicability of the durably iterated Prisoner's dilemma is much less restricted than it may first appear and extended his results with biologist William Hamilton to draw numerous conclusions about biological cooperation in terms of when, why, and how it does and does not appear. Most important is his observation that as one moves up the evolutionary ladder in neural complexity, behavior becomes richer because the intelligence of primates, including humans, allows a number of relevant improvements.

Precursors to altruism 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 acceptance 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.

Frans de Waal points out (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.***

Omohundro's behavioral predictions turn out to be the AI version of the obsolete "Veneer Theory" (Waal 2006) that "views morality as a cultural overlay, a thin veneer hiding an otherwise selfish and brutish nature". Before the effectiveness of morality was explained by game theory and logic, this was the dominant approach to morality within evolutionary biology and among writers popularizing the field. That field has advanced and now agrees that cooperative or social behavior is a more effective long-term strategy than being visibly selfish and is indeed evolutionarily selected for.

The Moral Sense

Human beings are clearly born with something that appears to be a sense of ethics, and develop that sense extensively as they grow to adulthood. In reality though, this "moral sense" is actually the result of biologically and socially selected "what's best for me, my gene pool or my society" process. We have evolved to feel good when we act altruistically because rational altruism is a survival trait. We have evolved to feel bad when we do something mean because acting unethically can be extremely detrimental to our survival and well being if or when we get caught. These feelings are reinforced by the fact that we can see that, generally (but not always), we reap what we sow.

If we could always use this "sense" and count on it to be accurate in determining an individual's and society's best interest, we wouldn't have many of the problems that we have in the world today and the path to safe AI would be much clearer. The problem is that, once the issues become non-trivial, the human moral sense is overwhelmed by evolved societal rules and norms in the general case, shortsighted/selfish motives in specific cases, and fear in both cases.

In addition, 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.

While this seems sub-optimal at first, this is a logical response to the evolution of our ability to sense deception in others. Keeping the reasoning inaccessible to the conscious mind allows the subconscious to make selfish judgments while making it easier for the conscious speaking mind to justify them (i.e. reverse a defection) without being hampered by fear of detection. Further, the unconscious mind can also alter the flow of the conscious mind's logic, frequently by focusing on just a few factors, without the conscious mind being aware of the fact (and the conscious mind has plenty of reason not to look further or more closely). 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.

Thus, self-justifying 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. Strong knee-jerk reactions need to be examined -- particularly where individuals utterly refuse to discuss or think about issues.

Altruism is an obvious case in point. As evidenced by the previously cited papers, evolution clearly selects for altruism once an organism evolves the necessary cognitive complexity to avoid having that altruism ruthlessly exploited. This, plus the evidence from game theory, should make it clear that altruism is in our self-interest. However, the fact that this is only most frequently the case, rather than an absolute logically derivable guarantee, means that many so-called "logical" individuals won't see why a very intelligent machine might make such a choice since their selfish subconscious won't let them.

A further problem is that, for the average human being, conscious logic is sub-optimal for making complex evaluations even when moral issues aren't involved. A study of the "deliberation-without-attention" effect (Dijksterhuis et al. 2006) shows clearly that engaging in a thorough conscious deliberation is only advantageous for simple choices while choices in complex matters should be left to unconscious thought. This effect is attributed to the fact that a person can pay conscious attention to only a limited amount of information at once, which can lead to a focus on just a few factors and the loss of the bigger picture.

The Hurdle of Universalism Over Relativism

Finally, there is also the critical problem that the moral sense appears to vary from culture to culture. The eminent developmental psychologist Lawrence Kohlberg (Kohlberg 1983) argued that our moral judgments are based on explicit rules and concepts and that virtually all of the fundamental components, rules, and judgments are universal. Inspired by the work of Jean Piaget and fascinated by children's reactions to moral dilemmas, he believed that over the course of a person's lifetime development, the rules and concepts that they use to solve moral problems unfold in a well-defined, universal sequence of six stages, each more adequate at responding to moral dilemmas than the last.

Kohlberg claimed that the process of moral development was principally concerned with justice and that its development continued throughout the lifespan. In the earliest stages of childhood, moral reasoning is based upon consequences. Teens and many adults are strongly influenced by the external authority of societal views and expectation. In later stages, moral reasoning appeals first to internalized convention, and then to general principles of neutrality, egalitarianism, and universal rights. To test this hypothesis, moral dilemmas were presented to people of varying ages and classes, both sexes, and many cultures. Kohlberg's key methodological insight was to focus not on the answers that people give to moral dilemmas but on how they justify their choice.

Kohlberg's claims were deeply controversial. The psychologist Carol Gilligan argued (Gilligan 1982) that women justify their moral choices differently from men, but with equal sophistication. Men, she claimed, tend to reason about morality in terms of justice, and women in terms of care: "While an ethic of justice proceeds from the premise of equality—that everyone should be treated the same—an ethic of care rests on the premise of non-violence—that no one should be hurt." Similar arguments were made for non-Western cultures—that they emphasize social roles and obligations rather than individual rights and justice.

The Linguistic Analogy

Fortunately, the most prominent feature of other cultures shows us the way over the hurdle. As mentioned repeatedly by Noam Chomsky but first detailed in depth by John Rawls (Rawls 1971), the study of morality is highly analogous to the study of language since we have an innate moral faculty with operative principles that cannot be expressed in much the same way we have an innate language faculty with the same attributes. Chomsky transformed the study of language and mind by claiming (Chomsky 1986) that human beings are endowed with an innate program for language acquisition and developing a series of questions and fundamental distinctions. Chomsky and the community of linguists working within this framework have provided us with an exceptionally clear and compelling model of how such a cognitive faculty can be studied.

Particularly relevant for the development of AI, is the fact that it is entirely likely that language is a mind-internal computational system that evolved for internal thought and planning and only later was co-opted for communication. Steven Pinker argues (Pinker 2007) that studying cross-cultural constants in language can provide insight into both our internal representation system and when we switch from one model to another. The fact that language dramatically affects our moral perceptions argues that they both use the same underlying computational system and that studying cross-cultural moral constants could not only answer what is moral but how we think and possibly even why we talk. De Waal (de Waal 2006) makes the interesting comment that primates spend an inordinate amount of time making up and we note that a brief, eloquent justification or an apology coupled with a recitation of why we value the relationship can turn a defection into a mistake or even an attempt at cooperation -- a huge evolutionary advantage.

Both language and morality are cognitive systems that can be characterized in terms of principles or rules that can construct or generate an unlimited number and variety of representations. Both can be viewed as being configurable by parameters that alter the behavior of the system without altering the system itself and a theory of moral cognition would greatly benefit from drawing on parts of the terminology and theoretical apparatus of Chomsky's Universal Grammar. The moral philosopher, like the linguist, needs to extract and state with full explicitness both these rules and the possible parameters. Marc Hauser (Hauser 2006) seems to have a very good program for studying the details of ethics by combining primate studies, infant studies, child development studies, brain scans, studies of unconscious actions, and cross-cultural studies of carefully designed novel situations ("trolley problems") but what hasn't been seen is a good over-arching top-down design.

Starting From The Top – The Societal View

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.

The opposite "top-down" approach is to start by defining the purpose (or goals) of ethics 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). Right and wrong can only be judged in light of a goal (or goals). Clearly, since evolution selects for them, one of the goals of ethics and altruism is survival.

Humans and other primates are generally ethical because society and other individuals demand it and punish deviance when it is caught. Animals that are sufficiently cognitively advanced are generally altruistic because the trickle-down effects of rational altruism, if not the direct rewards, generally end up well in excess of its costs. Thus, evolution favors ethics and altruism because, in the location of state space that humans and other cognitively advanced (and as a result, social) animals occupy, they promote self-interest and survival.

For the implementation of safe AI, the important questions are "Can we extend these observations to discover and implement a universal ethical system that follows the 5 S's (**S**imple, **S**afe, **S**table, **S**elf-correcting and **S**ensitive to current human thinking, intuition, and feelings)" and "Does this state space include or can it be extended to include super-intelligent machines?" Arguably, there are four fundamental reasons why a rational intelligence would not follow a society's dictates regardless of what they were: over-riding self-protection (fear); selfishness (greed); unfairness (error) on society's part; or error on the entity's part. If we can cover these four points, the path to safe AI should be clear.

It is important to note that, at this point, we still have no definition of what ethics is except that it is what is demanded by society. Yet, nothing more is required. Since, inarguably, the goal of a society is solely to fulfill the goals of its individual members, what an intelligent society is going to demand, as ethics, is nothing more or less than what is required to most optimally fulfill the goals of its constituent entities.

This provides the bedrock of the foundation of ethics, a universal yardstick by which the morality of an action can be judged. Actions that effectively further the goal(s) of society should be judged as ethically correct or "right". Actions contrary to the goals of society should be judged "wrong". Determining, under incomplete knowledge and uncertainty, what best furthers the goal(s) of society can be incredibly complex but this is finally a simple solid starting point for addressing that problem. Most importantly, it shows that ethics only exists in the context of a society. We are going to need to convince our super-intelligent AIs that it is in their own self-interest to join ours.

So, Why Shouldn't I Just Harvest You?

As De Garis points out, future super-intelligences may have the same relationship to our entire society that our current society has to an individual human – that of being able to control it or destroy it at will. The notion of societally enforced ethics just isn't going to apply. The good news, however, is that despite appearances, ethics didn't evolve from the top down because society suddenly magically enforced them.

Ethics evolved from the bottom up, one relationship at a time, because it is relationships that are beneficial to survival, self-interest, and goal-fulfillment. What an intelligent society enforces is merely what an intelligent non-shortsighted individual should be practicing for his own self-interest. Unfortunately, human beings are still unintelligent enough to be easily distracted by the lure of short-term personal payoffs at the cost of negative sum effects on society as a whole – the aggregate effect of which then trickles down to create more negative utility for each and every human being than the vast majority gained by "cheating".

The Relationship View

Instead of phrasing the foundation of ethics in terms of society's demands, it is more accurate and comprehensive to say that the foundation of ethics is relationships and society's demands are merely the incomplete and sometimes inaccurate wisdom compiled by evolution. In this view, ethics should be defined as what is best for the number and quality of relationships that we have. Since virtually any goal can be more easily achieved with friends and certainly without other entities deliberately interfering, establishing and maintaining such relationships is the sine qua non of truly intelligent behavior – and hopefully the AI becomes smart enough to see this becoming powerful enough to be beyond the control of society.

When the basis of morality is relationship, the measure of the morality of an action can be best regarded as a spectrum running from intelligent altruism to stupid selfishness. Intelligent altruism is incurring some personal negative utility solely in order to increase the sum of the perceived utility for all the entities in a relationship. If you are in relationship with entities that have an equal commitment to the relationship and intelligent altruism, everyone will benefit tremendously. Intelligent altruism 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. On the other hand, 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, morality must also be judged that way with each entity judging their own utility. This deferral of utility to the affected entity immediately allows and gives rise to the relativity of the cultural variations in behavior that have been problematical in the past.

Fairness, 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.

The most important 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.

This is not to say that we shouldn't apply utilitarian principles to promote ethics. Cheating is a defection and punitive/retaliatory action is the required response dictated by game theory and why it shows up in the moral sense and primate behavior. One entity with bad intentions means skyrocketing costs, as numerous measures need to be taken to ensure personal safety in the event of their defection. In fact, the intelligent view states that not only is restitution fair and that the cost of the process should be included but that a large enough punitive aspect should always be included so that entities will stop making the utilitarian calculation as to whether it is worth attempting to cheat.

Preparing for Take-off

Monomaniacal machines with a single anti-social goals like filling the universe with paperclips will see no value in a relationship with us and will have no reason not to harm us and will **always** be a danger. We must avoid creating such situations and ensuring that machines always have relationship goals is an easy and ethical solution.

Also, while we have protection from an ethical AI since all actions that involve us must be justifiable to us, it's an "as you sow, so shall you reap" type of situation. For example, if we believe that we are justified in creating an AI that is a slave to humanity's will, then we cannot complain if we are enslaved in turn. Yudkowsky's "safe" goals proposal (Yudkowsky 2004) is just such a relationship and should make any ethical person want to alter the AI's circumstances. It is also not "self-correcting" in the event of error or outside interference.

Correctly predicting the take-off of a truly moral AI is unlikely but we can make some very reasonable guesses. One likely guess is that the AI would want to "smarten" its society – which is likely to include an attempt to resolve the inequities that currently exist among humans. Even doing this by persuasion instead of force and attempting to minimizing the total amount of trauma as much as possible, there still is room for an awful lot of transition trauma given the amount of trauma currently induced upon the less fortunate.

We probably want to clean up our act before this happens – and we'll be much happier if we do even if AI never appears. One way in which to do this and to move towards the creation of AI in the process would be to create a collaborative system to discover (in the sense of Chomsky's framework) morality and our internal structures of thought and helps us move forward to a more intelligent (morally acceptable) world state.

References

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

Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York, NY: Praeger Publishers.

Dijksterhuis, A.; Bos, M.; Nordgren, L.; and Baaren, R. van 2006 On Making the Right Choice: The Deliberation-Without-Attention Effect. *Science* 311: 1005 – 1007.

de Garis, H. 2005. *The Artilect War: Cosmists Vs. Terrans.* Palm Springs, CA: ETC Publications.

Gilligan, C. 1982. *In a Different Voice*. Harvard Univ. Press.

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.

Kohlberg, L; Levine, C.; and Hewer, A. 1983. *Moral Stages: A Current Formulation and a Response to Critics*. Basel, Switzerland: S. Karger AG.

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

Omohundro, S. M. 2008a. The Basic AI Drives. In *Proceedings of the First Conference on Artificial General Intelligence*, 483-492. Amsterdam: IOS Press.

Omohundro, S. M. 2008b. *The Nature of Self-Improving Artificial Intelligence*. Available at http://selfawaresystems.files.wordpress.com

Pinker, S. 2007. The Stuff of Thought: Language as a Window into Human Nature. New York, NY: Viking/Penguin Group.

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.

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

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

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

Yudkowsky, E. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. Available at http://singinst.org/CFAI.html.

Yudkowsky, E. 2004. *Coherent Extrapolated Volition*. Available at http://www.singinst.org/upload/CEV.html.

Yudkowsky, E. 2006. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Bostrom, N. and Cirkovic, M. eds. *Global Catastrophic Risks*. Forthcoming.