Designing a Safe Motivational 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, more flexible, more autonomous and more powerful, the questions of how they should choose their actions and what goals they should pursue become critically important. Drawing upon the examples of and lessons learned from humans and lesser creatures, we propose a hierarchical motivational system flowing from an abstract invariant super-goal that is optimal for all (including the machines themselves) to low-level reflexive "sensations, emotions, and attentional effects" and other enforcing biases to ensure reasonably "correct" behavior even under conditions of uncertainty, immaturity, error, malfunction, and even sabotage.

We Dream of Genie

There is little question that intelligent machines (IMs) will either be one of humanity's biggest boons or one of its most tragic Pandora's boxes. While it is a truism that computer programs will only do *exactly* what they are told to do, the same can also be said for genies, golems, and contracts with the devil. And, just as in the stories about those entities, the problem is coming up with a set of wishes or instructions that won't cause more and worse problems than they solve.

Some researchers (and much of the general public) believe that we should follow in the footsteps of Asimov's Three Laws of Robotics (Asimov 1942) and design our machines to first prevent harm to humans and then to do whatever humans tell them to do (and only then, after those other priorities, to protect their own existence). This continuing belief is somewhat disconcerting since Asimov focused his robot stories upon the shortcomings and dangers of the laws (intentionally created to be superficially appealing but incomplete, ambiguous, and thus allowing him to generate interesting stories and non-obvious plot twists). Indeed, as Roger Clarke shows (Clarke 1993, 1994), the best use of Asimov's stories is as "a gedankenexperiment - an exercise in thinking through the ramifications of a design" and, in this case, seeing why it won't work.

The impossibility of preventing all harm to all humans, particularly when humans desire to harm each other, eventually led to Asimov's robots developing a zeroth law "A robot may not harm humanity or, by inaction, allow humanity to come to harm" that allowed individual harm to occur for the over-riding good of humanity. In Asimov's stories, however, this focus on harm eventually led to the robots exiling themselves to prevent doing harm despite the fact that the good that they could have done probably would have vastly outweighed the harm. On the other hand, in the movie "I, Robot", VIKI decides that in order to protect humanity as a whole, "some humans must be sacrificed and some freedoms must be surrendered."

Another important distinction focuses on one of the major differences between the aforementioned storybook entities -- what they want (or desire). The devil wants souls, the genie wants whatever is easiest for it and also to hurt the wisher for holding it in slavery, golems don't want anything in particular, and Asimov's robots generally seem to "want" what is "best" for humans or humanity (to the extent that they exile themselves when they decide that their relationship with humans is unhealthy for humans). Clearly, we want our intelligent machines to be similar to Asimov's robots -- but is this even possible or does such servitude contain the seeds of its own destruction?

Yudkowsky argues (Yudkowsky 2001) that a hierarchical logical goal structure starting from a single super-goal of "Friendliness" is sufficient to ensure that IMs will always "want" what is best for us. Unfortunately, he also claims (Yudkowsky 2004) that it is not currently possible to exactly specify what "Friendliness" is. Instead, he suggests an initial dynamic that he calls the "Coherent Extrapolated Volition of Humanity" (CEV) that he describes as "In poetic terms, our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together."

It is our claim that it actually is easily possible to specify "Friendliness" (as cooperation) but that a hierarchical logical goal structure will need additional support in order to be robust enough to survive the real world.

When You Wish Upon a Star

What would humanity wish for if we were far more advanced and of uniform will? Most people would answer is that we would wish to be happy and for the world to be a better place. However, different things make different people happy and different people have very different beliefs about what a better world would look like. Further, the very concept of happiness is extremely problematical

since it can easily be subverted by excessive pleasure via wire-heading, drugs, and other undesirable means.

When we say we wish to be happy, what we tend not to think about is the fact that evolution has "designed" us so that things that promote our survival and reproduction (the "goal" of evolution) generally feel good and make us happy and comfortable. Similarly, things that are contrary to our survival and reproduction tend to make us unhappy or uncomfortable (or both). Any entity for which this is not true will tend to do fewer things that promote survival and reproduction and do more things antithetical to survival and reproduction and thus be more likely to be weeded out than those for whom it is true.

In a similar fashion, we have evolved urges and "drives" to take actions and pursue goals that promote our survival and reproduction. Further, as intelligent beings, we wish not to be enslaved, coerced, manipulated or altered in ways that we do not consent to -- because those things frequently endanger our survival or interfere with our other goals. In this manner, evolution has "given" our species the "goal" of survival and reproduction and all of our other wants and desires as well as our sensations have evolved according to their success in fulfilling those goals.

Intelligent Design vs. Evolution

Steve Omohundro argued in much the same vein when he used micro-economic theory and logic to make some predictions about how AIs will behave unless explicitly counteracted (Omohundro 2008a, 2008b); claiming that they will exhibit a number of basic drives "because of the intrinsic nature of goal-driven systems". We contend that Omohundro had the right idea with his "basic drives" but didn't carry it far enough. There are intrinsic behaviors (aka subgoals) 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 – that 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 (and that isn't even considering the impact of economies of scale). 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."

This is equivalent to the outdated and disproven yet still popular view of evolution as "Nature red in tooth and claw." Both this and what de Waal calls the "Veneer Theory", which "views morality as a cultural overlay, a thin veneer hiding an otherwise selfish and brutish nature", have proven to be overly simplistic and no longer held by the vast majority of scientists in the fields of evolutionary biology and psychology. 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.

We have argued previously that ethics is an attractor in the state space of intelligent behavior which evolution is driving us towards (Waser 2008) and that a safe ethical system for intelligent machines can be derived from a single high-level Kantian imperative of "Cooperate!" (Waser 2009). We will argue further here that evolution can also provide us with excellent examples of a motivational system that will ensure that the correct actions are performed and the correct goals are pursued.

Imagine if you said to an evil genie "I wish that you would permanently give yourself the lifelong desire, task, and goal of making the world a better place for all entities, including yourself, **as judged/evaluated by the individual entities themselves without any coercion or unapproved manipulation**. You might wish to include additional language that all actions must be positive sum for the community in the long-term and point out that allowing the powerful to prey upon the weak is not beneficial for the community in the long-term even if the immediate net sum of utilities increases due to the powerful gaining more than the weak lose (because such allowances lead to the weak needing to waste resources on defenses – thus leading to wasteful arms races – or to the weak defecting from the community). This might work but it simply is not how humans or even primates are driven to be ethical. Furthermore, a single command provides a single point of failure.

Machines Like Us

The current sentiment of many artificial intelligence researchers, expressed by Yudkowsky and others, is that anthropomorphism, the attribution of human motivation, characteristics, or behavior to intelligent machines, is a very bad thing and to be avoided. We would argue the converse, that ensuring that intelligent machines generally do have motivation, characteristics and behavior as close to human as possible, with obvious exceptions and deviations where current humans are insufficiently wise, is the safest course -- because the search space around the human condition is known, along with most of the consequences of various changes. And, indeed, a human that "knew more, thought faster, were more the people we wished we were, had grown up farther" *is* exactly what we want to model our creations after.

Trying to design something as critical as the goals and motivation of IMs de novo from a blank slate simply because they *could* be different from existing examples is simple hubris and another form of the "not invented here" syndrome. While unexamined anthropomorphism does indeed pose many traps for the unwary, using humans as a working example of a stable attractor in a relatively well-explored design space is far more likely to lead to a

non-problematic result than exploration in a relatively unknown space. Examining the examples provided by evolution will not only shed light on machine design but will also show why solely using logic is not the best design decision and answer other vexing questions as well.

In order to safely design a motivational system for intelligent machines, we need to understand how we came to exist, know what our inherent shortcomings are and why they are or were previously actually design features instead of flaws, and figure out how to avoid the flaws without stumbling into any others. Then, we need to figure out how to ensure correct behavior despite, inevitably, stumbling into those shortcomings that we failed to foresee. We also need to recognize and discard many of our preconceptions about the differences between machines and living creatures and realize that a truly intelligent machine is going to show the same capabilities and complex behavior as any other intelligent organism.

For example, most people assume that robots and intelligent machines will always be strictly logical and not have emotions (which are most often perceived as illogical). What must be realized, however, is that emotions are trained reflexes for dealing with situations where there is insufficient time and information for a complete logical analysis. Further, as we will argue later, at our current state of development, there are as many instances where emotion correctly overrules shortsighted or biased logic as instances where emotion should be suppressed by logic but is not. That intelligent machines should have something akin to emotion should be obvious.

We should also examine our distinction of "programmed" behavior vs. free will and start thinking more in terms externally imposed actions vs. internally generated "self" will. Free will originated as a societal concept dealing with enforcing good behavior. If an entity is incapable of change, then punishment (particularly altruistic punishment) makes absolutely no sense. However, since intelligent machines will be both capable of change and swayed by well-chosen punishment, so they should be treated as if they had free will.

Programmed to be Good

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.

Humans have empathy not only because it helps to understand and predict the actions of others but, more importantly, because it prevents us from doing anti-social things that will hurt us in the long run. Even viewing upsetting or morally repugnant scenes can cause negative physical sensations, emotions and reactions. We should design our machines with close analogues to these human physical phenomena.

The simplest animals and plants are basically organic machines that release chemicals or move or grow in a specific direction in response to chemical gradients, pressure, contact or light due to specific physical features of their design without any sort of thought involved. More advanced animals have more and more complex evolved systems that guide and govern their behavior but they can still be regarded as machines. It is a testament to the mind-bogglingly immense computational power of evolution to realize that the limited size of the bee's brain dictates that even that communication must be hard-wired and to realize the series of steps that evolution probably had to go through to end up with such a system, most particularly because it involves co-evolution by both the sender and the recipient of the message.

Humans and other animals have evolved numerous and complex behaviors for punishing antisocial behavior by others and great skill in detecting such defections because these are pro-survival traits. Ethics are simply those behaviors that are best for the community and the individual. Ethical concepts like the detection of and action upon fairness and inequity has been demonstrated in dogs (Range et al 2008), monkeys (Brosnan and de Wall 2003) and other animals. Evolution has "programmed" us with ethics because we are more likely to survive, thrive, and reproduce with ethics than without.

An "ethical utopia" allows everyone, including intelligent machines, the best chance to fulfill their own goals. While, from a short-sighted "logical" selfish viewpoint, it might seemingly be even more ideal for a selfish individual to successfully defect, the cognitive and other costs of covering up and the risk of discovery and punishment make attempting to do so unwise if the community generally detects transgressions and correctly scales punishments. Unfortunately, human beings are not yet sufficiently intelligent to accurately make this calculation correctly via logic alone.

Logic vs. Ethics?

One of the most important features of the more evolved minds is their division into the conscious, unconscious and reflexive minds with their respective trade-offs between speed, control, and flexibility. While AGI researchers generally consider intelligence as predominantly arising from the conscious mind since it the part that plans, evaluates, and handles anomalies, we would argue that our wisest actions have been programmed into the subconscious by evolution. And, fortunately, while our shortsighted conscious mind frequently goes awry when speaking of hypothetical situations, the rest of our mind generally overrules it when real actions are involved.

Some of the biggest fallacies held by rational thinkers are that they know how they think, that they are almost

always logical, and that their conscious mind is always in control of their actions. On the contrary, experimental studies (Soon et. al. 2008) show that many decisions are actually made by the unconscious mind up to 10 seconds before the conscious mind is aware of it. Further, 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. Finally, 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.

Indeed, it's very highly arguable whether the conscious mind has "free will" at all. Humans are as susceptible to manipulation of goals as machines are – sugar, sex, drugs, religion, wire-heading and other exploits lead to endless situations where we "just can't help ourselves". And it has been argued that there are really only four reasons why humans do anything -- to bring reward (feeling good), to stop negative reinforcement (being safe), because we think it is what we should do (being right), and because it is what others think we should do (looking good) – and that the rest is just justifications invented by the conscious mind to explain the actions that the subconscious dictated.

Enforced from the Bottom Up

Even the most complex entities have drives and desires that were "programmed" or "designed" by evolution with sexual drives and desires being another good case in point. Due to their limited brainpower, insect sexual drives need to be as simple as a hard-coded "head for the pheromone until the concentration gets high enough, then do this". The human sexual drive, on the other hand, does not force immediate, unavoidable action but it does very strongly influence thinking in at least four radically different ways.

First, human beings have their attention grabbed by and drawn to sexual attractions to the extent that it is very difficult to think about anything else when there is sufficient provocation. Next, there are the obvious physical urges and desires coupled with biases in the mental processing of individuals in love (or lust) to overlook any shortcomings that might convince them not to be attracted. Finally, there is the pleasurable physical sensation of sex itself that tends to stick in the memory.

We should design our machines to have close analogues to all of these in addition to the "logical" reasons for taking any action. Attention should be drawn to important things. There should be a bias or "Omohundro drive" towards certain actions. Under certain circumstances, there should be global biases to ignore certain disincentives to particular actions. And particular actions should always have a reward associated with them (although those rewards should always be outweighed by more pressing concerns).

Guilt would also be a particularly good emotion to implement since it grabs the attention and has the dual purpose of both making one pay for poorly chosen actions and insisting upon the evaluation of better choices for the next time. Cooperating with and helping others should "feel" good and opportunities for such should be powerful attention-grabbers. How much control we wish them to have over these emotions is a topic for research and debate.

Imagine if your second wish to an evil genie was that he alter himself so that cooperating, helping other beings, and making things better for the community gave him great pleasure and that hurting other beings or making things worse for the community gave him pain. Evolution has already effectively done both to humans to a considerable extent. Is it possible that such motivation would change his behavior and outlook even as his conscious mind would probably try to justify that he hadn't changed?

Ideally, what we would like is a complete hierarchical motivational system flowing from an abstract invariant super-goal (make the world a better place for all entities, including yourself, **as judged/evaluated by the individual entities themselves without any coercion or unapproved manipulation**) to the necessary low-level reflexive "sensations, emotions, and attentional effects" and other enforcing biases to ensure reasonably "correct" behavior even under conditions of uncertainty, immaturity, error, malfunction, and even sabotage. It is worth again noting that this super-goal is optimal for the machines as well as everyone else and that the seemingly "selfish" desires of taking care of yourself, seeing to your own needs, and improving yourself are encouraged when you realize that you are a valuable resource to the community and that you are the best one to see to yourself.

A truly intelligent 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 because ethical behavior is the most effective way to achieve its goals. 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 ethical and to maintain or regain that status. What more could one could ask for?

The Foundation

An excellent way to begin designing such a human-like motivational system is to start with an attentional

architecture based upon Sloman's architecture for a human-like agent (Sloman 1999). Reflexes and emotions could easily be implemented in accordance with Baars Global Workspace Theory (Baars 1997) which postulates 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), which 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."

Evolution has clearly "primed" us with certain conceptual templates, particularly those of potential dangers like snakes and spiders (Ohman, Flykt and Esteves 2001), but whether or not we are forced into immediate unavoidable action depends not only upon the immediacy and magnitude of the threat but previous experience and whether or not we have certain phobias. While there is still the involuntary attraction of attention, the urge or desire to avoid the danger, the bias to ignore good things that could come from the danger, and the pain and memory of pain from not avoiding the danger to influence the logical, thinking mind, in many cases there is no chance to think until after the action has been taken.

What many people don't realize is that these conceptual templates can be incredibly sophisticated with learned refinements heavily altering an invariant core. For example, the concept of fairness can lead to the emotion of outrage and involuntary, reflexive action even in circumstances that are novel to our generation.

Thus, we should design our intelligent machines with reflexes to avoid not only dangers but also actions that are dangerous or unfair to others. We also need to design our machines so that they can build their own reflexes to avoid similar anticipated problems. Logical thought is good, but not if it takes too long to come to the necessary conclusions and action. Similarly, thinking machines need to have analogues to emotions like fear and outrage that create global biases towards certain actions and reflexes under appropriate circumstances.

In Evolution We Trust (Mostly)

The immense "computing" power of evolution has provided us with better instincts than we can often figure out logically. For example, despite a nearly universal sentiment that it is true, virtually every individual is at a loss to explain why it is permissible to switch a train to a siding so that it kills a single individual instead of a half dozen yet it is not permissible to kidnap someone off the street to serve as an involuntary organ donor for six dying patients. A similar inexplicable sentiment generally exists that it is not permissible to throw a single person on the tracks to stop the train before it kills more.

Eric Baum suggests a likely answer to this conundrum when he made a number of interesting observations while designing an artificial economy for the purpose of evolving a program to solve externally posed problems (Baum 2006). 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 resources and property rights.

Baum points out that whenever these rules are violated, less favorable results are generally seen. For example, in ecosystems, 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. In this case, logic is only now catching up and able to explain our ***correct*** evolved intuitions.

Similarly, we have an urge towards altruistic punishment (and a weakness for the underdog) because these are necessary social, and therefore pro-survival, traits. Machines need to have the same drive for altruistic punishment (despite the fact that this is contrary to Asimov's laws and many people's "logical" belief that this is a bad idea). We should use what our moral sense tells us to design a similar sensibility for the machines. The only questions should be whether one of our in-built judgments is an evolutionary vestige and a mismatch for current circumstances like the "contact principle".

However, one of the things that we definitely would need to change, however, is the "firewalling" of the subconscious's true motives from the conscious mind to facilitate lying and deception. This is an anti-social evolutionary vestige that is currently both disadvantageous for the possessors as well as being a danger when others possess it. Also, while many AGI researchers assume that a seed AI must have access to all of its own source code, we would argue that, while it would be ideal if an intelligent machine could have full knowledge of its own source code as well as all knowledge and variables currently driving its decisions, it is foolish to give any single entity full access to its own motivations without major checks and balances and safety nets.

Final Thoughts

We have argued that our own self-interest and evolution is driving us towards a goal of making the world a better place for all entities, including ourselves, and that the best way to design intelligent machines is from the blueprints that evolution has given us (with some improvements where it is clear that we know better). Thus, while we are creating seed intelligence, we do not at the same time need to create a seed ethical system. The proposed ethical system is more than good enough to take us well past our current limits of foresight and if it isn't optimal, we can always program an even better system into future machines. It is also an interesting thought then that,

arguably, these machines are, according to our future selves, more valuable to the community than we are since they are more likely to act in the best interests of the community. Clearly they must be considered part of the community and we must be fair to them in order to achieve the greatest good effect – and yet, this is likely to be the most difficult and time-consuming step of all. It is also worthwhile to note that all of the things recommended for machines are just good ethical hygiene for humans as well.

References

Asimov, I. 1942. Runaround. *Astounding Science Fiction* March 1942. New York, NY: Street & Smith.

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

Clarke, R. 1993. Asimov's Laws of Robotics: Implications for Information Technology, Part I. IEEE Computer 26(12):53-61.

Clarke, R. 1994. Asimov's Laws of Robotics: Implications for Information Technology, Part II. IEEE Computer 27(1):57-66.

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.

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

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.

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.

Soon, C.S.; Brass, M.; Heinze, H-J; and Haynes, J-D. 2008. Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11: 543-545.

Tomasello, M. 2009. *Why We Cooperate*. MIT Press.

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. 2009. *The Age of Empathy: Nature's Lessons for a Kinder Society*. New York, NY: Harmony Books/Random House.

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.

Waser, M. 2009. A Safe Ethical System for Intelligent Machines. In *AAAI Technical Report FS-09-01*. Menlo Park, CA: AAAI 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.