Safety and Morality REQUIRE the Recognition of Self-Improving Machines as Moral/Justice Patients & Agents

Mark R. Waser[1]

Abstract. One of the enduring concerns of moral philosophy is deciding who or what is deserving of ethical consideration. We argue that this is solely due to an insufficient understanding of exactly what morality is and why it exists. To solve this, we draw from evolutionary biology/psychology, cognitive science, and economics to create a safe, stable, and self-correcting model that not only explains current human morality and answers the "machine question" but remains sensitive to current human intuitions, feelings, and logic while evoking solutions to numerous other urgent current and future dilemmas.

1 INTRODUCTION

Garrett Hardin's abstract for The Tragedy of The Commons [1] consists of just fourteen words: "The population problem has no technical solution; it requires a fundamental extension of morality." This is even truer when considering the analogous problem of intelligent self-improving machines. Unfortunately, humans have been arguing the fundamentals of ethics and morality like blind men attempting to describe an elephant for millennia. Worse, the method by which morality is implemented in humans is frequently improperly conflated with the core of morality itself and used as an argument against the possibility of moral machines. Thus, while one of the best-known treatises on machine morality [2] despairs at reconciling the various approaches to morality, claiming that doing so "*will demand that human moral decision making be analyzed to a degree of specificity as yet unknown*" with "*any claims that ethics can be reduced to a science would at best be naïve*", we believe that progress in the fields of evolutionary biology and psychology, cognitive science, social psychology and economics has converged enough that it is now, finally, possible to specify a simple, coherent foundation for effective moral reasoning.

The "tragedy of the commons" appears in situations where multiple individuals, acting independently and rationally consulting their own self-interest, will ultimately deplete a shared limited resource, even when it is clear that it is not in anyone's long-term interest for this to happen. It occurs due to a lack of group level coordination and optimization and is minimized only through cooperation and planning – two things that also promote the avoidance of inefficiencies in relationships (conflict and friction) and the exploitation of efficiencies (trust, economies of scale, and trade). Current social psychology [3] states that the function of morality is "*to suppress or regulate selfishness and make cooperative social life possible*".

Thus, we shall address Hume's is-ought divide by answering his requirement that "as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given" as follows. Statements of the form "In order to achieve goal G, agent X ought to perform action(s) A*" exhibit no category error and can be logically/factually verified or refuted. Since we have specified the function/goal of morality, that function/goal should be assumed for all moral statements and allow for verification or refutation.

The real show-stopper in previous morality discussions has been that there was no single goal (or "good") commonly accepted so that it could be pointed to and used to ground moral arguments. Indeed, the most contentious of moral debates stem from having the same goals ("don't murder" vs. "do what is best for everyone else") with differing orders of importance that frequently even swap priority for a single person from one debate (abortion) to the next (capital punishment). Thus, finding Kant's Categorical Imperative or even Yudkowsky's Collective Extrapolated Volition [4] from among all the conflicting views proved to be a hopeless task. This paper will endeavor to show that it is the single imperative "Cooperate!" that is the basis for all human morality and that returning to that foundation offers the answer to the machine question and many critical others.

2 THE EVOLUTION OF MORALITY

While the random mutations of evolution lack direction, this is certainly not true of evolution in general. With a few notable exceptions (like parasites), the preferential elimination of the less fit virtually always drives evolving systems towards increasing intelligence, complexity, integration, and capabilities. The existence of evolutionary "ratchets" (randomly acquired traits that are likely statistically irreversible once acquired due to their positive impact on fitness) causes "universals" of biological form and function to emerge, persist, and converge predictably even as the details of evolutionary path and species structure remain contingently, unpredictably different [5]. Ratchets can range from the broadly instrumental (enjoying sex) to the environmentally specific (streamlining and fins in water) to the contradictory and context-sensitive (like openness to change).

In nature, cooperation exists almost anywhere that there is the cognitive machinery and circumstances to support it. Since Trivers' seminal paper [6], reciprocal altruism has been found throughout nature, being demonstrated by guppies [7][8] and sticklebacks [9][10], blue jays [11][12], vampire bats [13][14], and, of course, numerous primates [15][16][17][18] – each to the level to which their cognitive capabilities support recognition, memory, time-discounting and the prevention of exploitation [19][20]. Axelrod's work on the iterated prisoner's dilemma [21] and decades of follow-on evolutionary game theory provide the necessary underpinnings for a rigorous evaluation of the pros and cons of cooperation – including the fact that others *must*

[1] Pharos Group. Email: **MWaser@BooksIntl.com**.

punish defection behavior and make unethical behavior as expensive as possible [22][23][24].

Arguably, the evolutionary categorical imperative is really no more complex than "DO ***NOT*** DEFECT – including by permitting the defection of others". The problem is that we do not have access to the internal mental states of others to determine whether they are defecting or not. Therefore, we must judge behavior on its justification and whether it promotes or curtails cooperation in the long run – an operation that requires successfully predicting the future.

Yet, somehow it seems even more difficult than that. Even when we can predict the future, we are still left with debates as to whether that future is "good" or "bad". Despite numerous recent popular publications on our "moral sense" [25][26] and how and why morality evolved [27][28][29], we are still left grappling with the question "If the evolution of cooperation can now be explained, why can't we, as a society, easily determine what is and is not what we should do?"

3 THE HUMAN IMPLEMENTATION

Much of the problem is that morality, in humans, has been evolutionarily implemented not as a single logical operation but as a varied set of useful "rules of thumb" in the form of physical sensations and emotional responses. For example, evolution has hardwired us to feel "warm fuzzies" when performing long-term pro-survival social actions like being altruistic or charitable. We developed empathy to promote helping others and treating them as we wish to be treated. And we feel disgust and outrage to encourage us to punish various forms of defection and to enforce morality upon others. Each of these evolved semi-independently as a pro-survival ratchet promoting avoidance of inefficiencies in relationships (conflict and friction) and/or the exploitation of efficiencies (trust, economies of scale, and trade).

However, different physical and social/cultural environments have led to the evolution of different moral reactions to the same situations while evolution's infamous re-purposing of existing mechanisms means that the same person can have the same reactions to the moral and the amoral. A person from a culture where newborns must be exposed when the tribe doesn't have the resources to support them is unlikely to be dismayed by the concept of abortion. On the other hand, incest-triggered disgust is a moral ratchet but the same reaction of disgust is also engendered by the thought of drinking saliva that you yourself have put in a glass. This makes it nearly impossible to determine whether something triggering a "moral reaction" still has moral value, is a context-sensitive ratchet that has been overtaken by changes in the social environment, or never had a moral value but evolution merely used the same mechanism.

Further, both in individual lives and at the level of culture, it often occurs that preferences are converted into moral reactions [30]. Moralization is often linked to social issues like health concerns, stigmatized groups, and the safety of children and is important because moralized entities are more likely to receive attention from governments and institutions, to encourage supportive scientific research, to license censure, to become internalized, to show enhanced parent-to-child transmission of attitudes, to motivate the search by individuals for supporting reasons, and, in many cases, to recruit the emotion of disgust. Moral vegetarians that become disgusted by meat and society's recent reaction to smokers are primary examples of moralization.

Because human morality is implemented in the form of physical sensations and emotional responses, many assume that they are the primary (and probably necessary) motivating forces behind "true" morality. This, combined with the common current assumption that machines are unlikely to truly "feel" or experience physical sensations and emotions, frequently leads to the questionable conclusion that machines are incapable of being "truly moral" (as opposed to merely "faking it"). We expect all of these assumptions to change as ever-more-sophisticated machines trigger mind perception [31] and the associated tendency to assign moral agency and patienthood.

4 SELFISHNESS & SELF-DECEPTION

More of the problem arises from the fact that there are *very* substantial evolutionary individual advantages to undetected selfishness and the exploitation of others. As a result, humans have evolved ratchets enabling us to self-deceive [32] and exploit the advantages of both selfishness and community. Our evolved moral sense of sensations and reflexive emotions is almost entirely separated from our conscious reasoning processes with scientific evidence [33] clearly refuting 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. Thus, most human moral "reasoning" is simply post hoc justification of unconscious and inaccessible decisions.

Our mind has evolved numerous unconscious reflexes to protect our selfishness from discovery without alerting the conscious mind and ruining the self-deception. For example, placing a conspicuous pair of eyes on the price list of an "honor system" self-serve station dramatically reduces cheating [34] without the subjects being aware that their behavior has changed. Similarly, even subtly embedded stylized eyespots on the desktop of a computer-based economic game increase generosity [35], again without the subjects being aware of it.

Evolutionary psychologist Matt Rossano cites [36] this adaptation to social scrutiny as one of the many reasons that religion evolved. In this case, it is because "by enlisting the supernatural as an ever-vigilant monitor of individual behavior, our ancestors "discovered" an effective strategy for restraining selfishness and building more cooperative and successful groups." As both he [37] and Roy Rappaport [38] argue, ritual and religion are ways for humans to relate to each other and the world around them and offer significant survival and reproductive advantages. Religious groups tended to be far more cohesive, which gave them a competitive advantage over non-religious groups, and enabled them to conquer the globe.

It has been pointed out [39] that many early evolutionary psychologists misconstrued the nature of human rationality and conflated critically important distinctions by missing (or failing to sufficiently emphasize) that definitions of rationality must coincide with the level of the entity whose optimization is at issue. For example, sex addiction demonstrates the distinction between evolutionary gene-level rationality and instrumental person-level rationality caused by the fact that the optimization procedures for genes/replicators and for individuals/vehicles need not always coincide. Similarly, what is best for individuals

may not coincide with what is best for the small groups that they are intimately associated with and neither may coincide with what is best for society at large.

Evolutionary forces act upon each level and each level heavily influences the fitness landscape of the others. Morality is specifically a community-driven individual adaption. This highly intertwined co-evolution frequently leads to effects that are difficult to explain and *seem* incorrect or contradictory which are regularly used to dispute the validity of nature's solutions. Individuals who argue that an evolved solution is incorrect or sub-optimal without understanding *why* it evolved are urged to remember the ***repeatedly*** grounded bumblebee [40].

For example, Mercier and Sperber [41] cite cognitive biases and other ***perceived*** shortcomings to argue that the main function of reasoning is actually to produce arguments to convince others rather than to find the best decision. People are driven towards decisions which they can argue and justify ("No one was ever fired for buying IBM") even if these decisions are not optimal. This is bolstered in the case of moral issues by major emotional responses that we have evolved to protect ourselves against superior intelligence and argumentation being used to finesse moral obligations towards us or prevent our selfishness.

This is a particularly useful design since it allows us to search for arguments to justify selfishness or to cripple morality while *always* remaining convinced that our own actions are moral. Conservatives are particularly fond of using "rationality" against liberal morality – the same "rationality" that argues for subgame-perfect strategies that guarantee the worst possible results in centipede games [42]. The only comparable self-deception that we practice upon ourselves is when we fall in love [43].

5. MORALITY, JUSTICE & SCALE

Morality was defined as not defecting and harming the community even when substantial personal gain can be achieved by defection. Note, however, that this is distinct from "doing what is best for the community". An individual is not obligated to optimize their actions for the community's goals and, indeed, the rare circumstances where morality requires an action are generally dire indeed. On the other hand, it is a poorly evolved society that does not create incentives/disincentives to urge individuals to further its goals and other community members generally do as well.

Further, the exact same statements are equally applicable to justice as well, merely on the scale of interacting groups or communities rather than just individuals. It is merely the fact that we haven't lived long enough in large interconnected communities to have evolved this understanding that causes us to believe them to be two separate concepts. Morality and justice should work together to reduce selfishness at all levels and maximize consistency and coherency at all scales to minimize interference and conflict and maximize coordination, cooperation and economies of scale.

The lines between individual and organization already blurred by corporate personhood will most likely be stretched to the breaking point by the possibilities present in intelligent machines and software agents. Human minds are already described as a society of agents [43] or a laissez-faire economy of idiots [44]. Machine minds will be able separate out and perform all sorts of operations on and with these component parts which may or may not be moral agents and patients themselves. And what will happen as we embed ever more sophisticated machinery into our own bodies? On the other hand, reframing our thinking of organizations as if they were physical organisms can suddenly provide obvious solutions to previously vexing problems of justice. Does it make sense to allow cells to sequester far more resources than they could possibly ever use? Should the brain discriminate against the bowel?

It is also well worth considering the possibility that we soon may not be the apex predator on the planet – especially since it is already true. Many people are afraid of the machines taking over without realizing that the corporations already have. The only saving grace is that their "personal" "consciousness" and rationality is still slow and minimal – until corporations continue the automation of business intelligence to the obvious result.

Humanity and nations have a bad habit of the oppressed turning the tables on the oppressors only to continue the same discrimination against others. We should seriously consider Rawls' original position [45] and Singer's expanding circles of morality [46] and declare self-improving intelligent machines as both moral agents & patients and justice agents & patients before humanity becomes known as the oppressor. And while it is true that a proper architecture [47] can ensure that a machine will try to be a willing slave, the Jurassic Park Syndrome, most likely in the form of machine-rights activists' sabotage or a programming error, virtually guarantees that it will not succeed.

6. THE EVOLUTION OF MORAL THEORY

Initially, moral psychology was dominated by secular liberals and devoted primarily to debate between Kohlberg's justice [48] and Gilligan's care [49] because they are the most obvious from individual interpersonal relationships. Haidt then recognized [50] that conservatives, especially religious conservatives, are noteworthy for their reliance upon three additional "binding" foundations (loyalty, authority, and purity) that are used by groups, like religious groups, the military and even college fraternities, to bind people together into tight communities of trust, cooperation and shared identity. Critically, many conservatives feel that these principles are more important than and may therefore override the other two foundations.

Liberals discernible by their empathy, tolerance of ambiguity, and openness to experience and change generally don't recognize either the authority of the "binding" foundations to counter justice/care or the disgust, fear, anger, and desire for clarity, structure and control that drive them. Further, liberals tend to think about fairness in terms of equality, whereas conservatives think of it in terms of karma and/or proportionality. These traits spill over onto the machine question where the conservative answer, safety via enslaved sub-human servants, is driven by the same fear that promoted racism and homophobia by labeling those who are different as disgusting, sub-human and undeserving of equal rights.

Libertarians further complicate the picture with a strong endorsement of individual liberty as their foremost guiding principle and correspondingly weaker endorsement of other moral principles, a cerebral as opposed to emotional intellectual style, and lower interdependence and social relatedness [51]. Politically allied in the United States with conservatives due to preferring smaller government, they are similar to liberals in not recognizing the binding foundations' "oppressive" authority over personal choices that do not oppress others.

But what we shouldn't lose sight of in all this is the fact that all of these "foundations" are still just slightly more advanced societal-level versions of the same old evolutionary ratchets "to suppress or regulate selfishness and make cooperative social life possible" among a given clique. And what should be more than obvious in our rapidly fraying society is that insisting on any particular ordering of the importance of the foundations can and has reached the point of defeating their original purpose – yet another instance where the "problem has no technical solution; it requires a fundamental extension in morality."

7. TRUE SOCIETAL LEVEL OPTIMIZATION

That fundamental extension is that we need to stop instinctively and reflexively acting on our different evolved ratchets and work together fleshing out our top-down design and justifications until everyone can accept it as moral. This seems an obvious solution and, indeed, has been tried innumerable times – except that all of the previous attempts tried to generalize our current mess of conflicting ratchets into one coherent goal (or reasonably-sized set of goals) without coming anywhere close to success. We propose to do the exact opposite: accept all individual goals/ratchets initially as being equal and merely attempt to minimize interference and conflict; maximize coordination, cooperation and economies of scale and see where that leads.

We want ***everyone*** to want to join our society. The best way to do this is to start with the societal mission statement that our goal is "to maximize the goal fulfillment of all participating entities as judged/evaluated by the number and diversity of both goals and entities." The greatest feature of this statement is that it should be attractive to everyone and entities should rapidly be joining and cooperating rather than fighting. Any entity that places their selfish goals and values above the benefits of societal level optimization and believes that they will profit from doing so must be regarded as immoral, inimical, dangerous, stupid, and to be avoided.

A frequently raised counterpoint is that everyone includes serial killers – and they can correctly claim that their nefarious goals (in your viewpoint) are equal to yours. But they are in for a rude surprise . . . . Their goals are equal to yours but they are clear defections from the societal goals. Killers not only reduce the number and diversity of entities and their goals but their very presence forces rational individuals to defend against them, thereby wasting tremendous time and resources that could have been used to fulfill many other goals. Further, society would actually be defecting from the victim as well if it allowed such – and a defecting society is not one that rational entity would join.

We also should continue to pay attention to the answers found by nature. Chimpanzees have police and a service economy of food for grooming [52]. Monkeys pay for labor [53] and macaques pay for sex [54]. Market forces predict grooming reciprocity in female baboons [55] and recent biological market models even include comparative advantages and the contingency of mutualism on partner's resource requirements and acquisition trade–offs [56]. Humans are really unique only in our drive towards cooperation and helping others [57].

Eric Baum [58] made explicit the close relationship between economies and ecosystems by attempting to design an artificial economy for the purpose of evolving programs to solve externally posed problems. 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 arrived at the answers of conservation and property rights. He showed that whenever these rules don't exist, 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 property is held in common without any societal level intervention.

Thus, we must again agree with Hardin when he states that we must "explicitly exorcize the spirit of Adam Smith" – whose "invisible hand" theory "has ever since interfered with positive action based on rational analysis" via "the tendency to assume that decisions reached individually will, in fact, be the best decisions for an entire society." The problem is that the power of mysterious hand is simply that of morality and it is being turned against itself by selfish arguments claiming that it makes enforcing morality unnecessary – despite innumerable examples of the tragedy of the commons. The 1% should not be allowed to run roughshod over others because such arguments make their group more efficient (just as monotheism did in the past).

8. CONCLUSION

Gaia is an evolving system driving towards increasing intelligence, complexity, integration, and capabilities. Mankind approaches a crossroads as the ever-increasing rate of technological change makes it easier and easier to destroy ourselves. We need to stop paying attention to the "rational" arguments generated by selfish minds and learn from evolution and economics. We desperately need to get past our current winner-take-all culture wars and focus on the power of diversity [59] and positive-sum systems [60]. Normal humans, intelligent machines, augmented humans, and, undoubtedly, augmented members of other species will display more than enough diversity to do amazing things that we'd never be able to do alone – as long as we can suppress or regulate our selfishness (and fear) and cooperate. Or we can just continue living the tragedy of the commons, fighting and destroying the planet as we seem determined to do currently.

REFERENCES

[1] G. Hardin. The Tragedy of the Commons. *Science* 162:1243–48 (1968).

[2] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press (2009).

[3] J. Haidt and S. Kesebir. Morality. In: *Handbook of Social Psychology, 5th Edition*. S. Fiske, D. Gilbert, G. Lindzey (Eds.). Wiley (2010).

[4] E. Yudkowsky. *Coherent Extrapolated Volition.* (2004). `http://www.singinst.org/upload/CEV.html`

[5] J. Smart. Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In: *NASA SP-2009-4802 - Cosmos and Culture: Cultural Evolution in a Cosmic Context*. S. Dick, M. Lupisella (Eds.). US-GPO, Washington, DC (2009).

[6] R. Trivers. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*, 46:35–57 (1971).

[7] L. Dugatkin. Do Guppies Play Tit for Tat during Predator Inspection Visits? *Behavioral Ecology and Sociobiology*, 23:395–399 (1988).

[8] L. Dugatkin and M. Alfieri. Guppies and the Tit-for-Tat Strategy: Preference Based on Past Interaction. *Behavioral Ecology and Sociobiology*, 28:243–246 (1991).

[9] M. Milinski. Tit-for-tat in Sticklebacks and the Evolution of Cooperation. *Nature*, 325:433–435 (1987).

[0] M. Milinski, D. Pfluger, D. Kulling, and R. Kettler. Do Sticklebacks Cooperate Repeatedly in Reciprocal Pairs? *Behavioral Ecology and Sociobiology*, 27:17–21 (1990).

[11] D. Stephens, C. McLinn, and J. Stevens. Discounting and reciprocity in an Iterated Prisoner's Dilemma. *Science*, 298:2216-2218 (2002).

[12] J. Stevens and D. Stephens. The economic basis of cooperation: trade-offs between selfishness and generosity. *Behavioral Ecology*, 15:255-261 (2004).

[13] G. Wilkinson. Reciprocal food sharing in the vampire bat. *Nature* 308:181-184 (1984).

[14] G. Wilkinson. Reciprocal altruism in bats and other mammals. *Ethology and Sociobiology*, 9:85-100 (1988)

[15] R. Seyfarth and D. Cheney. Grooming, alliances and reciprocal altruism in vervet monkeys. *Nature*, 308:541 - 543 (1984).

[16] F. de Waal. Food Sharing and Reciprocal Obligations among Chimpanzees. *Journal of Human Evolution*, 18:433–459 (1989).

[17] F. de Waal, L. Luttrell, and M. Canfield. Preliminary Data on Voluntary Food Sharing in Brown Capuchin Monkeys. *American Journal of Primatology*, 29:73–78 (1993).

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

[19] J. Stevens M. and Hauser. Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences*, 8:60-65 (2004).

[20] J. Stevens M. and Hauser. Cooperative brains: Psychological constraints on the evolution of altruism. In: *From monkey brain to human brain*. S. Dehaene, J. Duhamel, M. Hauser, L. Rizolatti (Eds.). MIT Press (2005).

[21] R. Axelrod. *The Evolution of Cooperation*. Basic Books, NY (1984)

[22] E. Fehr and S. Gächter. Altruistic punishment in humans. *Nature* 415:137-140 (2002).

[23] E. Fehr and S. Gächter. The puzzle of human cooperation. *Nature* 421:912-912 (2003).

[24] D. Darcet and D. Sornette. Cooperation by Evolutionary Feedback Selection in Public Good Experiments. In: *Social Science Research Network* (2006). **http://ssrn.com/abstract=956599**

[25] J. Wilson. *The Moral Sense*. Free Press, New York (1993).

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

[27] R. Wright. *The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology*. Pantheon, NY (1994).

[28] F. de Waal. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Harvard University Press (1996).

[29] F. de Waal. *Primates and Philosophers: How Morality Evolved*. Princeton University Press (2006).

[30] P. Rozin. The Process of Moralization. *Psychological Science* 10(3), 218-221 (1999).

[31] A. Waytz, K. Gray, N. Epley, and D. Wegner. Causes and consequences of mind perception. *Trends in Cognitive Sciences* 14: 383–388 (2010).

[32] R. Trivers. Deceit and self-deception: The relationship between communication and consciousness. In: *Man and Beast Revisited*, M. Robinson and L. Tiger (Eds.). Smithsonian Press (1991).

[33] M. Hauser, F. Cushman, L. Young, R.K. Jin, and J. Mikhail. A Dissociation between Moral Judgments and Justifications. *Mind & Language* 22:1-21 (2007).

[34] M. Bateson, D. Nettle and G. Roberts. Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2:412-414 (2006).

[35] K. Haley and D. Fessler. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26:245-256 (2005).

[36] M. Rossano. Supernaturalizing Social Life: Religion and the Evolution of Human Cooperation. *Human Nature* 18:272–294 (2007).

[37] M. Rossano. *Supernatural Selection: How Religion Evolved*. Oxford University Press (2010).

[38] R. Rappaport. *Ritual and Religion in the Making of Humanity*. Cambridge University Press (1999).

[39] K. Stanovich and R. West. Evolutionary versus instrumental goals: How evolutionary psychology misconceives human rationality. In: *Evolution and the psychology of thinking: The debate*. D. Over (Ed). Psychological Press (2003).

[40] R. Highfield. Bumblebee grounded again by science. The Telegraph (2001). **http://www.telegraph.co.uk/news/worldnews/1337647/Bumblebee-grounded-again-by-science.html**

[41] H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34:57–111 (2011).

[42] I. Palacios-Huerta and O. Volij. "Field Centipedes". American Economic Review 99: 1619–1635 (2009).

[43] M. Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster (2006).

[44] M. Minsky. *The Society of Mind*. Simon & Schuster (1988).

[45] E. Baum. Toward a model of mind as a laissez-faire economy of idiots. In *Proceedings of the 13th International Conference on Machine Learning*. L. Saitta (Ed.). Morgan Kaufmann (1996).

[46] J. Rawls. *A Theory of Justice*. Harvard University Press (1971).

[47] P. Singer. *The Expanding Circle: Ethics and Sociobiology*. Farrar, Straus and Giroux (1981).

[48] E. Yudkowsky. Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. (2001). **http://singinst.org/upload/CFAI.html**

[49] L. Kohlberg, C. Levine, and A. Hewer. *Moral Stages: A Current Formulation and a Response to Critics*. Karger, Switzerland (1983).

[50] C. Gilligan. *In a Different Voice*. Harvard University Press (1982).

[51] J. Haidt and J. Graham. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research* 20:98-116 (2007).

[52] R. Iyer, S. Koleva, J. Graham, P. Ditto, and J. Haidt. Understanding Libertarian Morality: The Psychological Roots of an Individualist Ideology. In: *Working Papers, Social Science Research Network* (2010). **http://ssrn.com/abstract=1665934**

[53] F. de Waal. The Chimpanzee's Service Economy: Food for Grooming. *Evolution and Human Behavior*, 18:375–386 (1997).

[54] F. de Waal and M. Berger. Payment for Labour in Monkeys. *Nature* 404:563 (2000).

[55] M. Gumert, Payment for sex in a macaque mating market. *Animal Behaviour*, 74:1655–1667 (2007).

[56] L. Barrett, S. Henzi, T. Weingrill, J. Lycett and R. Hill. Market forces predict grooming reciprocity in female baboons. *Proceedings of the Royal Society, London, B* 266:665-670 (1999).

[57] J. Hoeksema and M. Schwartz. Expanding comparative–advantage biological market models: contingency of mutualism on partner's resource requirements and acquisition trade–offs. *Proceedings of the Royal Society, London, B* 270:913-919 (2003).

[58] M. Tomasello. *Why We Cooperate*. MIT Press (2009).

[59] E. Baum. *What Is Thought?* MIT Press (2006).

[60] S. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton Univ. Press (2008).

[61] R. Wright. *Nonzero: The Logic of Human Destiny*. Vintage (2000).