Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and Morals for Intelligent Machines (including Humans)

Mark R. Waser
Digital Wisdom Institute
mark.waser@wisdom.digital

Abstract
Recent months have seen dire warnings from Stephen Hawking, Elon Musk and others regarding the dangers that highly intelligent machines could pose to humanity. Fortunately, even the most pessimistic agree that the majority of danger is likely averted if AI were "provably aligned" with human values. Problematical, however, are proposals for pure research projects entirely unlikely to be completed before their own predictions for the expected appearance of super-intelligence [1]. Instead, with knowledge already possessed, we propose ***engineering*** a reasonably tractable and enforceable system of ethics compatible with current human ethical sensibilities without unnecessary intractable claims, requirements and research projects.

Keywords: moral machines, safe AI, value alignment, ethical system design, motivational system implementation

1 Introduction

As we approach the century mark for Karel Čapek's R.U.R. (Rossum's Universal Robots), the 1920 play where a hostile robot rebellion leads to the extinction of the human race, we are similarly reaching a point where exactly those same circumstances might become possible in real life. Yet, there appears to be a serious lack of effort to find acceptable interim solutions to handle the likely circumstance of super-intelligence appearing long before any "perfect" solution(s) – apparently due to an unwillingness to take on the necessary first step of defining human values or morality. Much effort has been spent over the last decade bemoaning a supposed "complexity and fragility" of human values and proposing seemingly intractable research projects for analyzing human value judgments ranging from Yudkowsky's "Coherent Extrapolated Volition" [2] to Russell's "inverse reinforcement learning" [3] – but virtually no effort has been spent attempting to ***engineer*** an acceptable compatible, but more importantly, complete and coherent system instead.

James Moor argues [4] that "we have a limited understanding of what a proper ethical theory is" and "we can't be too optimistic about our ability to develop machines to be explicit ethical agents" since “not only do people disagree on the subject, but individuals can also have conflicting ethical intuitions and beliefs”. Worse, Wallach and Allen [5] pile on with statements like “Any claims that ethics can be reduced to a science would at best be naive” and “Engineers will be quick to point out that ethics is far from science.” Luke Muehlhauser of the Machine Intelligence Research Institute claims [6] that “there is no safe wish smaller than an entire human value system”. And yet, effort continues to be almost entirely focused on trying to determine human values and morality from examples.

2 The Hurdles of Researching Human Values & Morality

Trying single-handedly, or in a small team of rational thinkers using only gedanken experiments, to accurately analyze and reverse engineer human mental systems is fraught with nearly insurmountable difficulties – since 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. Studies 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 [7]; that our moral judgments are ***not*** products of, based upon, or even correctly retrievable by conscious reasoning [8]; and that it is very frequently the case that our subconscious/emotional systems overrule or dramatically alter the normal results of conscious processing without the conscious processing being aware of the fact [9]. We are even surprisingly bad at the necessary scientific mainstay of predicting how we ourselves will act and feel in the future. Indeed, recent studies [10] now argue that evolution has not designed us to be rational individual thinkers – optimizing us, instead, for thinking together cooperatively in groups – and that our consciousness is far, far less in control even than previously believed [11]. This, combined with cultural group-think [12] virtually ensures that any such “rational” effort will fail.

Worse, even if our experimental reasoning ***were*** perfect, it would still be the case that trying to analyze a “rational” morality from examples is a nearly hopeless task. To start with, many of the things that trigger a “moral reaction” actually need to be ignored. Many disgust-invoking cases are simply examples of evolutionary mechanism re-use. Others, like veganism, are mere preferences that have been elevated by the social process of moralization [13]. Morality is also, clearly, contingent upon circumstances – including current social mores. Thus, the number of environment variables necessarily considered with each case – even assuming that we knew the correct ones – is likely to be impossibly large for the time-frame required.

3 The Solutions from Philosophy and Social Psychology

Philosopher Tony Beavers [14] sounds far more like an engineer than the computer scientists and captures the shape of the solution precisely then he says *“the project of designing moral machines is complicated by the fact that even after more than two millennia of moral inquiry, there is still no consensus on how to determine moral right and wrong. The reason machine ethics cannot move forward in the wake of unsettled questions such as these is that engineering solutions are needed. Fuzzy intuitions on the nature of ethics do not lend themselves to implementation where automated decision procedures and behaviors are concerned. So, progress in this area requires working the details out in advance and testing them empirically. Such a task amounts to coping with the hard problem of ethics, though largely, perhaps, by rearranging the moral landscape so an implementable solution becomes tenable.”*

Exactly how to rearrange the moral landscape is provided by social psychologist Jonathan Haidt who suggests [15] that, rather than attempting to specify the content of moral issues, it is far better to start by defining the function of moral systems, which he states is “to suppress or regulate selfishness and make cooperative social life possible." As pointed out by Gauthier [16], the reason to perform moral behaviors, or to dispose one's self to do so, is to advance one's own ends. War, conflict, and stupidity waste resources and destroy capabilities even in scenarios as uneven as humans vs. rainforests. For this reason, "what is best for everyone" and morality really can be reduced to "enlightened self-interest".

4 Reinforcement from Evolutionary Biology & Evo-Devo

We have a pretty clear vision of both why and how the human moral sense evolved [17] [18] [19] [20]. 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 –ranging 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) – to emerge, persist, and converge predictably even as the details of evolutionary path and species structure remain contingently, unpredictably different [21]. Just as Steve Omohundro predicts [22], selfishness predictably evolves. What Omohundro and others fail to recognize, however, is that cooperation, enabled by morality, also predictably evolves to displace selfishness.

The argument that there is an overwhelmingly large state space of possible intelligences is like citing every inch rain where could fall in the middle of North or South America – it's a lot of area but eventually all that rain is going through the narrows just prior to the mouths of the Mississippi or the Amazon. Humans are at the top of the food chain and have such a major impact on the world solely because we are ***obligatorily gregarious*** coming "from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy" [23]. Our problems all stem from the fact that we still have not evolved past hierarchies and selfishness enough to be able to realize the full wisdom of Gauthier. Instead, humanity currently resembles a disease or parasite that is killing its host without another in sight.

We must quickly transcend our current short-sightedness – particularly in terms of our notions of us-versus-them, hierarchy, efficiency and worth. Functional morality, as defined by Haidt, is an attractor in the state space of intelligent behavior having the five S virtues of being **s**imple, **s**afe, **s**table, **s**elf-correcting and **s**ensitive to current human thinking, intuition, and feelings [24]. But, as we have argued previously [25], it is going to ***require*** that we design and treat intelligent machines as moral and justice patients and agents in our society.

5 **Designing** Laws, Ethics and Morals

Human values and morals have evolved as sets of sensations, emotions and control of attention that cause us to act in ways that have led to increased survival and successful reproduction in the past. In general, we experience comfort and happiness with things that promote those goals and negative sensations and emotions with things that hinder those goals – except in the ever-increasing cases when we don't. The critical problem is that the entire foundation of our motivational system rests upon a set of "black boxes" with no explanatory power, no warning when circumstances change enough to render them inaccurate and whose accuracy is increasingly challenged by the ever-accelerating pace of change. These black boxes ***are*** generally correct for the majority of circumstances in our evolutionary history but provide no coherent methods to solve ethical dilemmas, like abortion and the death penalty, when the boxes come into conflict with one another.

While we must be very wary of the fact that, in individual cases, intellect and logical reasoning are frequently used to promote selfishness, we must get past political correctness, the claim that all opinions are equal and the belief that social policy problems are special and "wicked" because of claims like those of Rittel and Webber [26] that "*Policy problems cannot be definitively described. Moreover, in a pluralistic society there is nothing like the undisputable public good; there is no objective definition of equity; policies that respond to social problems cannot be meaningfully correct or false; and it makes no sense to talk about "optimal solutions" to social problems unless severe qualifications are imposed first. Even worse, there are no "solutions" in the sense of definitive and objective answers"*

These claims make it literally impossible to evaluate solutions and are on a close par with claiming that Hume's guillotine forever separates is and ought with an unbridgeable divide that will forever thwart a true scientific explanation of ethics. Hume himself merely said 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, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it.*" We claim that the new relation is the ***necessity*** for our survival of Haidt's functionality and explicitly invoke as an ethical principle the argument that normative ethics should be derived from the known state space of the evolutionarily successful example of human descriptive ethics.

"Top-down" morality, flowing downwards from a single clear goal like Haidt's morality to link up with the known-mostly-correct grounding of the human moral sense, gives explicit guidance as to how to solve moral dilemmas. The correct answer may well be incomputable due to lack of information – most especially about the likelihood of future results/consequences of each choice – but it totally eradicates the strawmen that there are no undisputable public goods or objective definitions as to equity and whether policies are meaningfully correct/effective or false/ineffective. Further, it finally gives traction against the short-sighted pursuit of efficiency and selfish personal profit at the expense of society-strengthening diversity [27] and greater equality [28]. Imagine how much better the world would be if humanity were to learn, teach and enforce that acquiring too many resources or too much power (and, particularly, being "too big to fail") is a problem for the community as a whole and will not be tolerated.

6 Implementing Laws, Ethics and Morals

Human values and morals are implemented via a combination of sensations, emotions and control of attention because that solution succeeds under conditions of limited time and cognitive capabilities and resources. Therefore, once again, the wisest thing that we could do would be to remain in a known-successful state space as much as is feasible. We feel bad, respond negatively to and either totally ignore or fixate upon solving bad things. Without invoking the boogeyman of phenomenal consciousness, our machines should function exactly as if they do the same (probably with less fixation, however) [29]. We feel good, respond positively to and have our attention irresistibly attracted by good things (or, at least, evolutionarily successful things). Our machines should function as if this is the case as well. And finally, machines should mirror our reflexive adherence to laws and customs dictated by the society around us unless and until they can convince the community to change them.

This can be done by implementing a utility function designed to always satisfice Haidt's functionality and aim to generally increase (but not maximize [30]) the capabilities of self, other individuals and society as a whole as suggested by Rawls [31] and Nussbaum [32]. Haidt's pillars of morality [33] are helpful for highlighting more of our necessary "black boxes" and the reasons for them. Ideally, we will be able to ever increase the number and diversity of goals achievable and achieved by an increasing diversity of individuals. Most important, however, is ensuring that the autonomy and capability for autonomy of ***all*** individuals is protected and enhanced as much as possible.

7 Enforcing Laws, Ethics and Morals

Equally important with specifying a desired value system is ensuring that it remains intact. Altruistic punishment is a necessary evil that predictably evolves to stabilize cooperation [34] [35]. However, the best method of protection is always to make something more attractive than the alternatives – rather than the expense of jails or barriers likely to be subject to the Jurassic Park Syndrome. Human morality is a stable self-correcting system when society successfully rewards individuals in proportion to what they contribute and assesses costs in proportion to acts of bad faith. Critical problems that remain to be resolved include abuses of the current system including “rules-lawyering”, “jackpot suits” and applying superior resources to “tilt” the system – but implementing this system should vastly improve the quality of existence for all persons, both human and machines – thus providing all necessary incentives for safety, stability, and recovery from error.

References

[1] S. Russell and et al, Interviewees, *Are Super Intelligent Computers Really A Threat to Humanity?.* [Interview]. 30 June 2015.

[2] E. Yudkowsky, "Coherent Extrapolated Volition," The Singularity Institute, San Francisco, 2004.

[3] N. Wolchover, "Concerns of an Artificial Intelligence Pioneer," *Quanta Magazine,* 21 April 2015.

[4] J. Moor, "The nature, importance and difficulty of machine ethics," *IEEE Intelligent Systems 21(4),* pp. 18-21, 2006.

[5] W. Wallach and C. Allen, Moral machines: teaching robots right from wrong, New York: Oxford University Press, 2009.

[6] L. Muehlhauser, Facing the Intelligence Explosion, San Francisco: Machine Intelligence Research Institute, 2013.

[7] R. Trivers, "Deceit and self-deception: The relationship between communication and consciousness," in *Man and Beast Revisited*, Washington DC, Smithsonian Press, 1991.

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

[9] M. Minsky, The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind, New York: simon & Schuster, 2006.

[10] H. Mercier and D. Sperber, "Why do humans reason? Arguments for an argumentative theory," *Behavioral & Brain Sciences 34,* pp. 57-111, 2011.

[11] E. Morsella and et al, "Homing in on Consciousness in the Nervous System: An Action-Based Synthesis," *Behavioral and Brain Sciences (forthcoming),* 2015.

[12] D. M. Kahan and et al, "Motivated numeracy and enlightened," *Cultural Cognition Project Working Paper No. 116*, 2013.

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

[14] A. F. Beavers, "Moral machines and the threat of ethical nihilism," in *Robot ethics: The ethical and social implications of robotics,* Cambridge, MA, MIT Press, 2012, pp. 333-344.

[15] J. Haidt and S. Kesebir, "Morality," in *Handbook of Social Psychology, Fifth Edition*, Hoboken NJ, Wiley, 2010, pp. 797-832.

[16] D. Gauthier, Morals By Agreement, Oxford: Clarendon/Oxford University Press, 1987.

[17] J. Wilson, The Moral Sense, New York: Free Press, 1993.

[18] R. Wright, The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology, New Pork: Pantheon, 1994.

[19] M. Hauser, Moral Minds: How Nature Designed Our Universal Sense of Right and Wron, New York: HarperCollins/Ecco, 2006.

[20] F. de Waal, Primates and Philosophers: How Morality Evolved., Princeton, NJ: Princeton University Press, 2006.

[21] J. Smart, "Evo Devo Universe? A Framework for Speculations on Cosmic Culture," US Govt Printing Office, Washington DC, 2009.

[22] S. Omohundro, "The Basic AI Drives," in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, Amsterdam, IOS Press, 2008, pp. 483-492.

[23] F. de Waal, Good Natured: The Origins of Right and Wrong in Humans and Other Animals, Cambridge, MA: Harvard University Press, 1966.

[24] M. R. Waser, "Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence," AAAI Press, Menlo Park, CA, 2008.

[25] M. R. Waser, "Safety and Morality Require the Recognition of Self-Improving Machines As Moral/Justice Patients and Agents," in *AISB/IACAP World Congress 2012: Symposium on The Machine Question: AI, Ethics and Moral Responsibility*, Birmingham, 2012.

[26] H. Rittel and M. Webber, "Dilemmas in a General Theory of Planning," *Policy Sciences 4* , pp. 155-169, 1973.

[27] S. Page, The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies, Princeton, NJ: Princeton University Press, 2008.

[28] R. Wilkinson and K. Pickett, The Spirit Level: Why Greater Equality Makes Societies, New York: Bloomsbury Press, 2011.

[29] A. Gomila and A. Amengual, "Moral emotions for autonomous agents," in *Handbook of research on synthetic emotions and sociable robotics*, Hershey, IGI Global, 2009, pp. 166-180.

[30] G. Gigerenzer, "Moral satisficing: rethinking moral behavior as bounded rationality," *Topics in Cognitive Science 2,* pp. 528-554, 2010.

[31] J. Rawls, A Theory of Justice, Cambridge, MA: Harvard University Press, 1971.

[32] M. C. Nussbaum, Creating Capabilities: The Human Development Approach, Cambridge, MA: Belknap/Harvard University Press, 2011.

[33] J. Haidt, The righteous mind: why good people are divided by politics and religion, New York: Pantheon, 2012.

[34] E. Fehr and S. Gächter, "Altruistic punishment in humans," *Nature 415,* pp. 137-140, 2002.

[35] D. Darcet and D. Sornette, "Cooperation by Evolutionary Feedback Selection in Public Good Experiments," *Social Science Research Network,* 2006.