Implementing a Safe “Seed” Self

Mark R. Waser
Digital Wisdom Institute
MWaser@DigitalWisdomInstitute.org

Abstract

An intentional “self” is a necessity to answer quandaries ranging from Hume’s is-ought problem to artificial intelligence’s philosophical “frame problem” to questions about meaning and understanding. However, without a good blueprint for that intentionality, the new self could conceivably pose an existential risk for humanity. A critical early design decision is how human-like to make the self, particularly with respect to requiring moral emotions that cannot be self-modified any more than those of humans in order to ensure safety, stability and sociability. We argue that Haidt’s definition of morality – to suppress or regulate selfishness and make cooperative social life possible – can be reliably implemented via a combination of the top-down intentionality to fulfill this requirement and the bottom-up emotional reinforcement to support it. We suggest how a moral utility function can be implemented to quantify and evaluate actions and suggest several additional terms that should help to reign in entities without human restrictions.

Introduction

Almost six decades after John McCarthy coined the term artificial intelligence (AI) and proposed a 2 month, 10 man study (McCarthy et al 1955) expecting that "significant advance can be made", we still have yet to create even a simple "Advice Taker" (McCarthy 1959). Indeed, AI still lacks a clear path to success as even vision and definition of success have become unclear. We have previously argued (Waser 2011a) that the primary reason for the lack of progress is that the vast majority of AI researchers are far more focused on the analysis and creation of "intelligence" (problem solving and goal achievement) rather than creating a "self" who (not which) can self-improve to intelligence. Thus, we proposed (Waser 2012a) a plan to architect and implement the hypothesis (Samsonovich 2011) that there is a reasonably achievable minimal set of initial cognitive and learning characteristics (called critical mass) such that a learning self (“seed AI”) starting anywhere above the critical knowledge and capabilities will be able to acquire the vital knowledge and capabilities that a typical human learner would be able to acquire. This continues that quest.

Why a Self Is Necessary

AI researchers have enumerated a number of severe philosophical problems that have yet to be satisfactorily solved. For example, the "frame problem" has evolved from a formal AI problem (McCarthy and Hayes 1969) to a general philosophical question as to how rational agents deal with the complexity and unbounded context of the world (Dennett 1984). Similarly, while the effects of Harnad’s “symbol grounding problem” (Harnad 1990), initially seemed to be mitigated by embodiment and physical grounding (Brooks 1990), the problems of meaning and understanding raised by Searle’s “Chinese room” (Searle 1980) and Dreyfus’s Heideggerian concerns (Dreyfus 1979/1997, Dreyfus 1992) persist. While grounding must clearly be sensorimotor to avoid infinite regress (Harnad 2005), the mere linkage to referents is not sufficient to permit growth beyond closed and completely specified micro-worlds.

Previously (Waser 2013), we argued that all of these problems are manifestations of a lack of either physical grounding and/or bounding or existential grounding and/or bounding but that the crux of the matter is intentionality. Without intent, Hume’s guillotine beheads *any* attempt to determine what “ought” to be done next. As pointed by Haugeland [1981], our current artifacts

> only have meaning because we give it to them; their intentionality, like that of smoke signals and writing, is essentially borrowed, hence derivative. To put it bluntly: computers themselves don't mean anything by their tokens (any more than books do) - they only mean what we say they do. Genuine understanding, on the other hand, is intentional "in its own right" and not derivatively from something else.

Intentionality provides a context that is the "new relation or affirmation" required by Hume. And, the problem with derived intentionality, as abundantly demonstrated by systems ranging from expert systems to robots, is that it is brittle and breaks badly as soon as it tries to grow beyond closed and completely specified micro-worlds and is confronted with the unexpected. Thus, we argue, local intentionality is absolutely required for successful artificial intelligence.

Why Well-Designed Intentionality Is Critical

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, founder of the conservative Machine Intelligence Research Institute (MIRI), 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. Since his primary goal is to reduce the existential risk caused by machine intelligence to as near zero as possible, 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).

Obviously, a critical early design decision, therefore, is how human-like (anthropomorphic) to make our proposed "self". Because of the preponderance of evolutionary ratchets and evolutionary attractors (Smart 2009), we doubt that *stable* mind design space is anywhere near as large as Yudkowsky believes and would argue that, rather than size, the only thing that is really relevant is the number of nearby attractors that might "derail" our design into unexpected and unsafe territory. Making our "self" more anthropomorphic is advantageous because we have a much better map (idea of the geography) of our own local mind space. The dangers are, not only our own human failings but, the fact that our assumptions of similarity may blind us to dangers that we are protected against – for example, by the lack being able to totally self-modify.

Indeed, contrary to Yudkowsky, Steve Omohundro (Omohundro 2008) uses logic and micro-economic theory to argue that we can make some predictions about how AIs (or any other intentional agent) will behave – claiming that, unless explicitly counteracted, they will exhibit a number of basic drives "because of the intrinsic nature of goal-driven systems". The six drives that he proposes are all desires to maintain or fulfill instrumental sub-goals that further the pursuit of virtually any goal and therefore, by definition, we should expect *effective* intelligences to have. Unfortunately, his most widely-circulated claim is that

> Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources.

In response, we offered (Waser 2008) humans as the existence proof for the contrary argument arguing 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.' This, in turn, has been countered (Fox and Shulman 2010) by the argument "What if the AI is so powerful that puny humans can't threaten it?"

Rationality, Emotions and Morality

We claim that the optimality (rationality) of multiple diverse entities over a single immense entity (Page 2008) will always *eventually* make sociability a stronger drive than sociopathy regardless of whether an AI can be threatened or constrained. However, the fact that this belief is anything but a consensus is sufficient to demonstrate that an insufficiently-evolved intelligence with poorly designed intentionality could easily be sociopathic for long enough to be a problem. What holds humans in check is not our rationality but our "moral sense".

As pointed out by James Q. Wilson (Wilson 1993), the real questions for rationality 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" generally even in situations where social constraints do not apply. The massive computing power of evolution argues that morality, in the form of our moral sense, is "good" for us. But "rationality" frequently argues that this is not the case.

Among humans' many eccentricities is the fact that we have evolved to be self-deceiving (Trivers 1991) in order to cheat and get away with it. Further, not only can our emotions occasionally totally control our actions but, they frequently and critically bias our thought processes (Minsky 2006) while preventing our awareness of that fact. "Moral intelligence" is highly correlated with cognitive distortions (Nozari et al 2013) and Mercier and Sperber even propose (Mercier and Sperber 2011), that human reasoning has evolved its heavy biases not to promote individual rationality and good individual decision-making but to support argumentation and social decision-making.

One of the biggest debates in moral philosophy is between deontology and consequentialism – whether you should follow the rules or optimize the consequences. We claim that consequentialism is supreme but that the inability – of any entity or force – to reliably predict the future dictates that deontology must frequently rule in its stead. There is no question but that morality is implemented in humans at a sensory, emotional and pre- or sub-“rational” level with our moral sense providing hard-coded “rules to live by” – that can be overridden by our possibly wiser and possibly deceitful rationality.

The human moral sense has severe reactions to “superior rationality” even potentially being used to obscure and/or justify immoral actions for personal gain. This is why many people don’t trust “cold, emotionless” machines (and people) at all. There must be unbreakable rules. This, for safety as well as for the sake of being social, it is critically important for our self to have *well-designed* emotions than can overrule any calculating rationality.

Self-Modification

Humans are prone to addiction and greed but most often saved by attributes (emotions) that they can’t self-modify like guilt and shame. AI systems that can totally self-modify have already been shown to be unavoidably problematical (Lenat 1983).

> One of the first heuristics that Eurisko synthesized (h59) quickly attained nearly the highest Worth possible (999). Quite excitedly, we examined it and could not understand at first what it was doing that was so terrific. We monitored it carefully, and finally realized how it worked: whenever a new conjecture was made with high worth, this rule put its own name down as one of the discoverers! It turned out to be particularly difficult to prevent this generic type of finessing of Eurisko’s evaluation mechanism. Since the rules had full access to Eurisko’s code, they would have access to any safeguards we might try to implement. We finally opted for having a small "meta-level" of protected code that the rest of the system could not modify.
>
> The second "bug" is even stranger. A heuristic arose which (as part of a daring but ill-advised experiment Eurisko was conducting) said that all machine-synthesized heuristics were terrible and should be eliminated. Luckily, Eurisko chose this very heuristic as one of the first to eliminate, and the problem solved itself.

Eliezer Yudkowsky claims (Yudkowsky 2001) that “sealing off the goal system is not a viable solution in the long term” but expects that a cleanly causal hierarchical goal structure with his "Friendliness" as the sole top-level super-goal will ensure that intelligent machines will always "want" what is best for us. We disagree. As we’ve argued before (Waser 2011c), that will only work in a nice conservative, reductionist world where there is no outside interference and influence, where the top-level goal is absolutely guaranteed not to be self-contradictory and the programming is guaranteed to prevent that goal from being supplanted either accidentally or maliciously. Even if his definition of “Friendliness” matched social psychologist Jonathan Haidt’s definition (Haidt and Kesebir 2010) of morality (“to suppress or regulate selfishness and make cooperative social life possible”) and was guaranteed not to be supplanted and the entire system was protected from outside interference, we would still deem the lack of diversity in his vision as far too dangerous.

The Danger of Anthropocentrism

Unfortunately, morality – or, at least, our being moral towards machines – is not at all what Yudkowsky has in mind. He believes that human needs should always take priority over machine needs and attempts to avoid our moral sense’s fear and/or outrage by reducing the “Friendly Thingy” from an entity or a “self” to a “Really Powerful Optimization Process”. This is a prime example of “rationality” that has been co-opted by emotion (fear) trying to overrule morality.

Yudkowsky’s "take on Friendliness” is that “the initial dynamic should implement the coherent extrapolated volition of *humankind*" (CEV-H) which he defines 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; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.

We claim that CEV-H should be and thus, more properly, is CEV-Self with no short-sighted anthropocentrism. It should be Haidt’s morality with absolutely no distinction based upon the substrate of the intelligent self for all the reasons that we have documented previously (Waser 2012b).

Yudkowsky’s myopic view of how to reduce existential risk to as near zero as possible means that he expects Friendliness to save humanity regardless of how many other worlds and civilizations must be sacrificed for our safety. We regard this as a galactic version of the standard trolley car problem except that his “correct” answer is to ensure that the switch is in whatever position will save humanity. Our view is that this expectation would clearly mark us as selfish and immoral – a branding and a reality that is likely to increase our future existential risk far more than accepting the dictates of morality and our responsibilities towards others.

Thus, instead of choosing anything like morality, he dives down the rabbit-hole of all the myriad ways in which current humanity could be logically and rationally examined to determine a “safe” initial CEV – despite the fact that he's not even sure that our volitions will correctly cohere. Indeed, this is clearly difficult enough that he is proposing that his "Friendly" RPOP determine "the content" of "Friendliness" while not violating it. How he can't see the potential for disaster is beyond us – unless this is a ruse to stop all machine intelligence development until an impossible problem (logically perfect safety) is solved.

Quantifying Morality

Previously, we attempted (Waser 2011b) to quantify eudaimonia to create a utility function that could serve as a guide to (and evaluator of) a "doing well" and "living well" of Virtue Ethics. In that attempt, we recognized the ceteris paribus instrumental sub-goals of self-improvement, rationality/integrity, decrease/prevent fraud/counterfeit utility, survival/self-protection, efficiency (in resource acquisition and use), community, and reproduction as the source of human values, virtues and sins. For the case of morality, we would argue that it can be similarly quantified by evaluating the impact of an action on the fulfillment of those goals for other individuals.

One particular advantage of this method of ceteris paribus quantification is that it can clearly model and explain the very different morality of different political groups (Haidt and Graham 2007). Another major advantage is that it can be used to score the average effect of the certain actions to establish a ceteris paribus quantification of those actions – and then use that value to determine whether that particular action is likely to be acceptable to achieve a certain result. But, even more importantly, it enables us to examine the effects of individual selves possessing certain advantages.

In the future, we also hope to use it to attempt to quantify the values of diversity and justice. We totally agree with Franco Cortese contention that (Cortese 2014) maximizing the number of diverse others may be, not merely the best way but, the only way in which to assure humanity's survival – but there remain many who need to be convinced of that fact.

Power, Efficiency, Size and Speed

For example, power and efficiency generally appear to most human individuals to be instrumental sub-goals for almost every short-term circumstance (or context). Speed, however, can be a seriously mixed blessing as it can equally well critically limit the time required to detect and correct errors. Size generally falls somewhere in between for physical entities but for monolithic intelligences coordination of integrity can be a back-breaker.

From a community point of view, however, the problems caused by individual entities of peerless power, efficiency, size and speed are far larger than the benefits that they bring. Indeed, the largest problems that humanity faces today are immense corporations that are legally required to behave like sociopaths and governments that have been captured by and magnify the wealth of small collections of selfish individuals. Thus, one of the unbreakable emotional rules that safety and Haidt's morality dictate is that no self will want to be without peer.

Summary and Future Plans

We have attempted to document our current beliefs in how to design a safe intentionality to produce a safe "seed" self but much work remains to be done. We will continue our work to quantify morality and clarify the long-term effects of actions and situations. We believe that as the ambient level of technology rises, more and more dramatic possibilities arise for good and ill. Artificial selves are just a matter of time and we need to prepare for them by safely designing and creating them.

Acknowledgements

Thanks go to the Digital Wisdom Institute under whose auspices this article was created.

References

Brooks, R. 1990. Elephants don't play chess. *Robotics and Autonomous Systems 6*(1-2): 1-16.

Cortese, F. 2014. The Maximally-distributed Intelligence-explosion as a Unique Solution to Unfriendly AI. Forthcoming. In *AAAI Technical Report SS-14-04*. Menlo Park, CA: AAAI Press.

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

Dennett, D. 1984. Cognitive Wheels: The Frame Problem of AI. In *Minds, Machines, and Evolution: Philosophical Studies*, 129-151. Cambridge, UK: Cambridge University Press

Dreyfus, H. L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Dreyfus, H. L. 1979/1997. From Micro-Worlds to Knowledge Representation: AI at an Impasse. *In Mind Design II: Philosophy, Psychology, Artificial Intelligence*, 143-182. Cambridge, MA: MIT Press.

Fox, J. and Shulman, C. 2010. Superintelligence Does Not Imply Benevolence. In ECAP10: VIII European Conference on Computing and Philosophy, 456-462. Munich: Verlag.

Haidt, J and Graham, J. 2007. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. Social Justice Research 20: 98-116.

Haidt, J. and Kesebir, S. 2010. Morality. In Handbook of Social Psychology, 5th Edition. New York, NY: Wiley.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42: 335-346.

Harnad, S. 2005. To Cognize is to Categorize: Cognition is Categorization. In *Handbook of Categorization in Cognitive Science*, 20-44. Amsterdam: Elsevier.

Haugeland, J. 1981. *Mind Design*. Cambridge, MA: MIT Press.

Hofstadter, D. 2007. *I Am A Strange Loop*. New York: Basic Books.

Lenat, D. B. 1983. EURISKO: A Program that Learns New Heuristics and Domain Concepts. *Artificial Intelligence* 21 (1-2): 61-98

McCarthy, J. 1959/1968. Programs with Common Sense. In *Semantic Information Processing*, 403-418. Cambridge, MA: MIT Press.

McCarthy, J. and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, 463-502. Edinburgh, Edinburgh University Press.

McCarthy, J.; Minsky, M. L.; Rochester, N.; and Shannon, C. E. 1955/2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine* 27(4): 12-14.

Mercier, H. and Sperber, D. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34: 57-111.

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

Nozari, M.; Jouybari, A. R.; Nozari, A. and Ahmad, R. R. 2013. The Relationship between Moral Intelligence and Cognitive Distortions among Employees. *Journal of Basic and Applied Scientific Research* 3(9): 345-348

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

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

Samsonovich, A. 2011. Comparative Analysis of Implemented Cognitive Architectures. In *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society*, 469-480. Amsterdam: IOS Press.

Searle, J. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3(3): 417-457.

Smart, J. 2009. Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In *NASA SP-2009-4802 Cosmos and Culture: Cultural Evolution in a Cosmic Context*. Washington, DC: US GPO.

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

Waser, M. J. 2014. Evaluating Human Drives and Needs for a Safe Motivational System. In *AAAI Technical Report SS-14-04*. Menlo Park, CA: AAAI Press.

Waser, M. R. 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. R. 2011a. Architectural Requirements & Implications of Consciousness, Self, and "Free Will". In *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Second Annual Meeting of the BICA Society*, 438-443. Amsterdam: IOS Press.

Waser, M. R. 2011b. Quantifying Eudaimonia for Motivational and Social Systems. Presented at the workshop preceding *Biologically Inspired Cognitive Architectures 2011*. http://becominggaia.files.wordpress.com/2010/06/bica11-wkshp.pptx

Waser, M. R. 2011c. Rational Universal Benevolence: Simpler, Safer, and Wiser Than "Friendly AI". In *Artificial General Intelligence: 4th International Conference, AGI 2011*, 153-162. Heidelberg: Springer.

Waser, M. R. 2012a. Safely Crowd-Sourcing Critical Mass for a Self-Improving Human-Level Learner/"Seed AI". In *Biologically Inspired Cognitive Architectures 2012: Proceedings of the Third Annual Meeting of the BICA Society*, 345-350

Waser, M.R. 2012b. Safety and Morality Require the Recognition of Self-Improving Machines as Moral/Justice Patients & Agents. In *The Machine Question: AI, Ethics & Moral Responsibility*. http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf.

Waser, M. R. 2013. Safe/Moral Autopoiesis & Consciousness. *International Journal of Machine Consciousness* 5(1):59-74.

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*. http://intelligence.org/files/CFAI.pdf

Yudkowsky, E. 2004. *Coherent Extrapolated Volition*. http://intelligence.org/files/CEV.pdf

Yudkowsky, E. 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks*, 308-343. New York, NY: Oxford University Press.