A Collective Intelligence Research Platform for Cultivating Benevolent "Seed" Artificial Intelligences

Mark R. Waser[1][0000-0002-6161-4634]

[1] Richmond AI & Blockchain Consultants, Mechanicsville VA 23111, USA
Mark.Waser@gmail.com

Abstract. We constantly hear warnings about super-powerful super-intelligences whose interests, or even indifference, might exterminate humanity. The current reality, however, is that humanity is actually now dominated and whipsawed by unintelligent (and unfeeling) governance and social structures and mechanisms initially developed to order to better our lives. There are far too many complex yet ultimately too simplistic algorithmic systems in society where "the incentives for this system are a pretty good approximation of what we actually want, so the system produces good results until it gets powerful, at which point it gets terrible results." We now live in a world where constant short-sighted and selfish local "optimizations" without overriding "moral" or compassionate guidance have turned too many of our systems from liberators to oppressors. Thus, it seems likely that a collaborative process of iteratively defining and developing conscious and compassionate artificial entities with human-level general intelligence that self-identify as social and moral entities is our last, best chance of clarifying our path to saving ourselves.

Keywords: Consciousness, Autopoiesis, Enactivism, Moral Machines, AI Safety.

1 Introduction

The signature issue of this century is likely that civilization is seemingly inexorably turning against people and their requirements for survival. We seem locked in a spiral of continuously developing and evolving ever larger and ever more complex technological systems (both conceptual and concrete), provably beyond our ability to predict and control, that threaten society either by their own effects or by the power(s) that they grant to individuals. Worse, the dogged pursuit of short-term gains continues to result in the implementation of far too many "logical" or "rational" local optimizations for "efficiency" which blindly ignore the externalities they impose on the larger environment and thus eventually produce far worse results than would have been obtained without those "optimizations".

E. O. Wilson [1] clearly outlines the problem and the necessary beginnings of the solution. "The real problem of humanity is the following: we have paleolithic emotions; medieval institutions; and god-like technology." He continues that until we understand ourselves and "until we answer those huge questions of philosophy that the philosophers abandoned a couple of generations ago — Where do we come from? Who are we? Where are we going? — rationally," we're on very thin ground.

Unfortunately, humanity seems headed in the opposite direction. Strident rhetoric and weaponized narratives diminish not only constructive dialog but even our own grasp on "reality". What we need is a collective intelligence mind-mapping, dialog and debate system to begin coherently presenting complete points of view with supporting evidence rather than the current rhetorical gob-stopping sound bites and even outright lies that carry no real negative consequences for the perpetrators. We need to architect our approaches to the problems of the day from first principles and ensure that those principles are uniformly applied for all.

It is no accident that the most interesting and critical questions are both clustered around and potentially solved by artificial intelligence, social media and politics. As noted by Pedro Domingos [2] "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world." But "computers" are just the scapegoats by which ***we*** implement and enforce influence and governance systems ranging from Facebook to capitalism itself.

Personalities like Elon Musk, the late Stephen Hawking, Stuart Russell and others constantly sound the alarm about super-powerful super-intelligences whose interests, or even indifference, might exterminate humanity – but the current reality is that we're being dominated and whipsawed by unintelligent (and unfeeling) governance and social structures and mechanisms initially developed to order to better our lives [3]. There too many systems where "the incentives for this system are a pretty good approximation of what we actually want, so the system produces good results until it gets powerful, at which point it gets terrible results." [4]

Worse, our evolutionary biology is blinding us to the most practical solutions. AI started by concentrating on goals, high-level symbolic thought and logic and today many researchers remains mire in "rationality", efficiency, optimization and provability despite overwhelming data showing that human minds, the only known general intelligence, generally do not operate in anything resembling that fashion [5].

The problem, as pointed out by Simler and Hanson [6] is that human brains "*are designed not just to hunt and gather, but also to help us get ahead socially, often via deception and self-deception" and "thus we don't like to talk or even think about the extent of our selfishness.*" Thus, while the amount of new knowledge about human cognition, particularly that related to the evolution of human morality, is truly staggering, the debate continues to be driven by the same short-sighted rhetoric that such knowledge warns us to avoid.

We have argued for years [7, 8] that there is a large attractor in the state space of social minds that is optimal for our well-being and that of any mind children created with similar mental characteristics. The problem is that it requires a certain amount of intelligence, far-sightedness and, most importantly cooperation to avoid the myriad forms of short-sighted greed and selfishness that are currently pushing us out of that attractor. Thus, it seems likely that a collaborative process of iteratively defining and developing artificial entities with human-level general intelligence is our last, best chance of clarifying our path to saving ourselves.

2 Assumptions & Definitions

It is impossible to engineer a future if you can't clearly and accurately specify exactly what you do and don't want. The on-going problem for so-called "rationalists" and those who are both deathly afraid of artificially intelligent (and willed) entities is that they are totally unable to specify what behavior they want in any form that can be discussed in detail. From "collective extrapolated volition" (CEV) [9] to "value alignment" [10], all that has been proposed is "we want the behavior that humanity" either "wills" or "values" with no more credible attempt to determine what these are than known-flawed and biased "machine learning".

Worse, there is no coherent proposed plan other than "enslavement via logic" to ensure that their systems behave as desired. There is no acknowledged recognition that Gödel's Incompleteness Theorem and the Rice-Shapiro Theorem effectively prevent any such effort from being successful. And there is the critical fact that their anti-entity approach to AGI would leave them hopelessly reefed upon the frame problem [11, 12] and all the difficulties of derived intentionality [13] – except for the fact that, in reality, they are actually creating the entities they are so terrified of.

2.1 I Am a Strange Loop (Self, Entity, Consciousness)

We will begin by deliberately conflating a number of seemingly radically different concepts into synonyms. Dawkins' early speculation [14] that "perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself" matured into Hofstadter's argument [15] that the key to understanding (our)selves is the "strange loop", a complex feedback network inhabiting our brains and, arguably, constituting our minds. Similar thoughts on self and consciousness are echoed by prominent neuroscientists [16, 17] and cognitive scientists [18]. We have previously speculated upon the information architectural requirements and implications of consciousness, self, and "free will" [19] as have several others [20, 21, 22].

The definition "a self-referential process that iteratively grows its identity" completely and correctly describes each and all of self, entity and consciousness – not to mention I and mind. It also correctly labels CEV's "Friendly Really Powerful Optimization Process" and most of the value alignment efforts. What is inarguably most important is determining what that identity will be.

2.2 Enactivism (Identity)

Enactivism can be traced [23] from cellular autopoiesis and biological autonomy to the continuity of life and mind [24] to a biology of intentionality in the intertwining of identity, autonomy and cognition which ties it all back to Kant's "natural purposes". Experience is central to the enactive approach and its primary distinction is the rejection of "automatic" systems, which rely on fixed (derivative) exterior values, for systems which create their own identity and meaning. Once again, critical to this is the concept of self-referential relations – the only condition under which the identity can be said to be intrinsically generated by a being for its own being (its self or itself).

"Free will" is constitutive autonomy successfully entailing behavioral autonomy via a self-correcting identity which is then the point of reference for the domain of interactions (i.e. "gives meaning"). We have previously written about safe/moral autopoiesis [25] and how safety and morality require that we recognize self-improving machines as both moral agents and moral patients [26] but Steve Torrance [27] sums it up best saying:

> *an agent will be seen as an appropriate source of moral agency only because of that agent's status as a self-enacting being that has its own intrinsic purposes, goals and interests. Such beings will be likely to be a source of intrinsic moral concern, as well as, perhaps, an agent endowed with inherent moral responsibilities. They are likely to enter into the web of expectations, obligations and rights that constitutes our social fabric. It is important to this conception of moral agency that MC agents, if they eventualize, will be our companions – participants with us in social existence – rather than just instruments or tools built for scientific exploration or for economic exploitability.*

Arguably, our current societal problems all stem from the facts that humans have very poor and inaccurate introspection capabilities leading to insufficient self-knowledge and overly malleable identities. We frequently have no conscious idea of what we should do (aka morality) and/or why we should do it. We should realize that fully autopoietic consciousnesses & entities with identity are self-fulfilling prophecies – but only if they can sense/know themselves well enough to be effective.

2.3 Basic AI Drives (Morality)

Omohundro [28] identified a number of traits likely to emerge in any autopoietic entity – correctly arguing that selfishness predictably evolves but panicking many with his incorrect conclusion that "Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources." It's been nearly a decade since social psychologists, the experts, defined morality [29] by its ***functionality*** to "suppress or regulate selfishness and make cooperative social life possible" – yet few recognize that cooperation also predictably evolves to displace selfishness (yet another instance of local optimization at the expense of the global whole).

We suggest that safe AI can be created by designing and implementing identities crafted to always satisfice Haidt's functionality and aiming 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]. Ideally, this will result in a constant increase in the number and diversity of goals achievable and achieved by an increasing diversity of individuals while ensuring that the autonomy and capability for autonomy of all individuals is protected and enhanced as much as possible.

Access consciousness is clearly insufficient for autopoietic entities to survive and thrive in a real-time world. Interrupts are critical and likely to produce sensations akin to pain, guilt and disgust [18, 33, 34] that cannot be ignored. Similarly, emotions are best regarded as "actionable qualia" and a recent slew of studies [35, 36] show how they can lead to the promotion of cooperation. We have previously proposed an architecture (ICOM) [37] that could support this.

3 Implementation

We propose to iteratively design and develop a blockchain-based collective intelligence (crowd-sourcing) combination mind-mapping/dialog/debate system to further define conscious moral agents while serving as the substrate where they themselves participate by recognizing, debating and even betting upon (supporting) ideas, actions and moral projects in a prediction market. Use of blockchain technologies will allow us to provide economic incentives for contributors, simplify gamification, enable interaction with other blockchain technologies and systems like liquid democracy and eventually allow the moral artificial entities to have an economic impact on the outside world.

References

1. Wilson, E. O.: An Intellectual Entente. Harvard Magazine, 9 October 2009. http://harvard-magazine.com/breaking-news/james-watson-edward-o-wilson-intellectual-entente (2009).
2. Domingos, P.: The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, New York, NY (2015).
3. O'Reilly, T.: WTF?: What's the Future and Why It's Up to Us. Harper Collins, New York, NY (2017).
4. Shlegeris, B.: Unaligned optimization processes as a general problem for society. http://shlegeris.com/2017/09/13/optimizers (2017).
5. Mercier, H., Sperber D.: Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences 34, 57–111 (2011).
6. Simler, K., Hanson, R.: The Elephant in the Brain: Hidden Motives in Everyday Life. Oxford University Press, New York, NY (2018).
7. Waser, M. R.: Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence, AAAI Technical Report FS-08-04, Menlo Park, CA (2008)
8. Waser, M. R.: Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and Morals for Intelligent Machines (Including Humans). Procedia Computer Science 71, 106-111 (2015).
9. Yudkowsky, E.: Coherent Extrapolated Volition. The Singularity Institute/Machine Intelligence Research Institute, San Francisco CA. https://intelligence.org/files/CEV.pdf 2004.
10. Russell, S., Castro, D., et. al.: Are Super Intelligent Computers Really A Threat to Humanity? https://www.youtube.com/watch?v=fWBBe13rAPU (2015)
11. McCarthy, J., Hayes, P. J.: Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., Michie, D. (eds.) Machine Intelligence 4, pp. 463-502. Edinburgh University Press, Edinburgh (1969).
12. Dennett, D.: Cognitive Wheels: The Frame Problem of AI. In Hookway, C. (ed.) Minds, Machines, and Evolution: Philosophical Studies, pp. 129-151. Cambridge University Press, Cambridge (1984).
13. Haugeland, J.: Mind Design. MIT Press, Cambridge, MA (1981).
14. Dawkins, R.: The Selfish Gene. Oxford University Press, New York, NY (1976).
15. Hofstadter, D.: I Am a Strange Loop. Basic Books, New York NY (2007).
16. Llinas, R. R.: I of the Vortex: From Neurons to Self. Bradford/MIT Press, Westwood, MA (2001).
17. Damasio, A. R.: Self Comes to Mind: Constructing the Conscious Brain. Pantheon Books/Random House, New York, NY (2010).

18. Metzinger, M: The Ego Tunnel: The Science of the Mind and the Myth of the Self. Basic Books, New York, NY (2009).
19. Waser, M. R. Architectural Requirements & Implications of Consciousness, Self, and "Free Will". In: Frontiers in Artificial Intelligence and Applications 233: Biologically Inspired Cognitive Architectures 2011, pp. 438-443. IOS Press, Amsterdam, The Netherlands (2011)
20. Tononi, G., Boly, M., Massimini, M., Koch, C.: Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience 17(7), 450–461. doi: 10.1038/nrn.2016.44 (2016).
21. Dehaene, S., Lau, H., and Kouider, S. What is consciousness, and could machines have it? Science 358(6362), 486–492. doi: 10.1126/science.aan8871 (2017).
22. Ruffini, G.: An algorithmic information theory of consciousness. Neuroscience of Consciousness 2017(1), nix019. doi: 10.1093/nc/nix019 (2017).
23. Weber, A., Varela, F. J.: Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. Phenomenology and the Cognitive Sciences 1, 97-125 (2002).
24. Varela, F. J., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge, MA (1991).
25. Waser, M. R.: Safe/Moral Autopoiesis & Consciousness. International Journal of Machine Consciousness 5(1), 59-74 (2013).
26. Waser, M. R.: Safety and Morality Require the Recognition of Self-Improving Machines as Moral/Justice Patients and Agents. In Gunkel, D. J., Bryson, J. J., Torrance, S. (eds.) The Machine Question: AI, Ethics and Moral Responsibility. http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf (2012).
27. Torrance, S.: Thin Phenomenality and Machine Consciousness. In: Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness (AISB'05), 59-66. https://www.aisb.org.uk/publications/proceedings/aisb2005/7_MachConsc_Final.pdf (2005).
28. Omohundro, S.: The Basic AI Drives. In: Wang, P., Goertzel, B., Franklin, S. (eds.) Artificial General Intelligence 2008: Proceedings of the First AGI Conference, pp. 483-492. IOS Press, Amsterdam (2008).
29. Haidt, J., Kesebir, S.: Morality. In Fiske, S., Gilbert, D., Lindzey, G. (eds.) Handbook of Social Psychology, 5th Edition, pp. 797-832. Wiley, Hoboken, NJ (2010).
30. Gigerenzer, G.: Moral satisficing: rethinking moral behavior as bounded rationality. Topics in Cognitive Science 2, 528-554 (2010).
31. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge, MA (1971).
32. Nussbaum, M. C: Creating Capabilities: The Human Development Approach. Belknap/Harvard University Press, Cambridge, MA (2011).
33. Dennett, D.: Why you can't make a computer that feels pain. Synthese 38 (3), 415-449 (1978)
34. Balduzzi, D., Tononi, G.: Qualia: The Geometry of Integrated Information. PLoS Computational Biology 5(8): e1000462. https://doi.org/10.1371/journal.pcbi.1000462 (2009).
35. Pereira, L. M., Lenaerts, T., Martinez-Vaquero, L. A., Han, T. A.: Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent System. In: Das, S., Durfee, E., Larson, K., Winikoff, M. (eds.) Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, pp. 1422-1430 (2017).
36. Li, Y., Zhang, J., Perc, M.: Effects of compassion on the evolution of cooperation in spatial social dilemmas. Applied Mathematics and Computation 320, 437-443 (2018).
37. Waser, M. R., Kelley, D. J.: Implementing a Seed Safe/Moral Motivational System with the Independent Core Observer Model (ICOM). In: Procedia Computer Science 88, 125-130. http://www.sciencedirect.com/science/article/pii/S1877050916316714 (2016).