Bootstrapping a Structured Self-Improving & Safe Autopoietic Self

Mark R. Waser
Digital Wisdom Institute, Vienna, VA, U.S.
mWaser@digitalWisdomInstitute.org

Abstract
After nearly sixty years of failing to program artificial intelligence (AI), it is now time to grow it using an enactive approach instead. Critically, however, we need to ensure that it matures with a "moral sense" that will ensure the safety and well-being of the human race. Implementing consciousness and conscience is the next step the way towards creating safe and cooperative machine entities.

Keywords: enactive AI, "seed AI", moral machines

1 Introduction

Almost sixty years after the Dartmouth Summer Research proposal [1], it may finally be possible "to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves." What initially seemed a relatively simple problem in symbolic logic turned into a nearly bottomless quagmire that prevented growth beyond closed and completely specified micro-worlds. The "frame problem" quickly grew from a formal AI problem [2] to the more general philosophical problem of how entities deal with the unbounded nature of context and complexity in the real world [3]. Similarly, Harnad's grounding problem [4] initially seemed mitigated by embodiment [5], but the problems of meaning and understanding [6][7][8][9] eventually shoaled on Kant's natural purposes and Aristotle's teleology. Intentionality seemed a solution but became mired when extrinsic [10][11], derived [12] or merely "as-if" intentionality [13].

While Rodney Brooks tried to use his subsumption architecture [14] "to build complete creatures rather than isolated cognitive simulators" [15], he had to abandon the approach saying [16]:

> *Perhaps it is the case that all the approaches to building intelligent systems are just completely off-base, and are doomed to fail. Why should we worry that this is so? Well, certainly it is the case that all biological systems are:*
>
> - *Much more robust to changed circumstances than our artificial systems.*
> - *Much quicker to learn or adapt than any of our machine learning algorithms*
> - *Behave in a way which just simply seems life-like in a way that our robots never do.*
>
> *Perhaps we have all missed some organizing principle of biological systems, or some general truth about them. Perhaps there is a way of looking at biological systems which will illuminate an inherent necessity in some aspect of the interactions of their parts that is completely missing from our artificial systems.*

Conscious human reasoning and problem-solving is undoubtedly predominantly symbolic but it is now equally clear that it is critically built upon foundations that we are only now recognizing and replicating. It is only by returning to biology and the human example – Varela's re-enchantment of the concrete [17] – that progress can be made.

2 Enactive Artificial Intelligence

The enactive paradigm was initially conceived as an embodied and phenomenologically informed alternative to mainstream cognitive science [18] that grew out of biological autonomy and autopoiesis, the minimal organization of living systems [19][20]. As explained by Froese and Di Paolo [21], the enactive approach consists of a core set of ideas, namely autonomy, sense-making, emergence, embodiment, and experience, which find novel applications in a diverse range of disciplines such as biology, phenomenology, artificial life, social science, robotics, psychology, and neuroscience. Defining autonomy as organizational closure and the self-reference inherent in the process of self-production applies equally well to biological systems (the immune system, the nervous system, single-cell and multi-cellular organisms), social systems and mechanical systems and allows the enaction of a meaningful world through identity constitution. Better yet, the enactive approach also has a lot to say about social interaction forming the dynamics constitutive of both individual agency [22][23] and social cognition [24] and thus runs the gamut "from cell to society and back again".

Traditional robots and AI programs remain composed of “an externally defined collection of components that we have merely chosen to designate as an ‘agent’ by convention” [21] with arbitrary choices distinguishing the system from the environment [25] and which “only have meaning because we give it to them” [12]. In contrast, an enactive system is organized in such a way that its activity is both the ‘cause and effect’ of its own autonomous organization with its activity depending upon organizational constraints, which are in turn regenerated by the activity itself. It has an essentially self-constituted identity, because its own generative activity demarks what is to count as part of the system and what belongs to the environment, and meaning and understanding are generated relative to that active identity. Indeed, Weber and Varela [26] have gone so far as to propose “a basic revision of the understanding of teleology in biological sciences” by “accepting that *organisms are subjects having purposes according to values encountered in the making of their living*” and thus have an intrinsic/immanent teleology arising from their biological autonomy and biological individuality.

Searle [6] can only be answered when a system is formed by deep causal connections with the environment intertwining identity and cognition [27]. Programmed-in knowledge and actions provide only the shallowest and most brittle referents for the symbol structures that the system must manipulate. While Brooks was trying to program a creature, evolutionary methods were allowing robots to learn instead. Law and Miikkulainen’s approach [28] “explicitly rejects built-in task-specific knowledge, works within a continuous (simulated) environment and leaves the entire structure of the processing machinery up to evolution”. They focused on “learning the relation between sensors and effectors” tabula rasa rather than manually adding ever-increasing knowledge into the system. Di Paolo [29][30][31][32] focused on “organismically-inspired” robotics and reproduced homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. Oudeyer [33][34][35] tackled autonomous mental development with “The Playground Experiment” while Demiris and Dearden [36] ran the gamut from motor-babbling to hierarchical learning by imitation.

3 Structure and scaffolding vs. tabula rasa

Unfortunately, current self-organizing robotics has perhaps focused far too much on a tabula rasa approach. As argued by Pinker [37], human beings start as anything but a blank slate. The massive computing power of evolution has “programmed in” all sorts of useful structures ranging from our attention being grabbed when we spot a snake [38] to our innate moral sense [39][40][41][42][43]. And while morality predictably evolves [44], selfishness predictably evolves first [45] – so it is clearly in our best interests to figure out how to reliably pass moral structures on to our mind children.

Similarly, self-organizing robotics has not ventured far into the necessary realm of lifelong learning [46][47]. We have developed many machines that have learned to move in an incredibly lifelike fashion [48] but they are primarily reactive with virtually no high-level control, prediction or on-line reasoning. Now is the time, therefore, that we should be talking about implementing functional consciousness on top of their effectively unconscious but robust learning capabilities. Sensor readings can be used to create a grounded virtual reality that such a consciousness could control and live in exactly as human selves do [49][50][51][52].

4 Consciousness and Conscience

Autopoiesis completes Hofstadter's Strange Loop [53], allows the cognitive self to come to the physical mind [54] and even gives traction on the hard problem of consciousness [55]. As suggested previously [56], we believe the best way to do this is with a blackboard operating system similar to Hofstadter's CopyCat [57] or LIDA [58] based upon Baars' Global Workspace model of consciousness [59]. Consciousness then runs as a process to detect anomalies, learn, and generally act like the Governing Board of the Policy Governance model [60] to create a consistent, coherent and integrated narrative plan of action to meet the goals of the larger self per Dennett's narrative model of self [61].

We would probably want to maintain some of the low-level evolved features of human consciousness like automatic subjective referral of the conscious experience backwards in time [62][63] while enhancing transparency wherever possible. For example, humans are far too prone to illusory agency with subliminal and supraliminal priming enhancing experienced authorship [64] and even inducing false illusory experiences of self-authorship [65][66]. We should not replicate the fact that our conscious, logical mind are constantly self-deceived to enable us to most effectively pursue what appears to be in our own self-interest [67][68]. And it would be particularly helpful if machine moral judgments were products of, based upon, and correctly retrievable by conscious reasoning – as opposed to the human case [69][70] – based upon the social psychologists' functional definition of morality [71] reinforced by sensory incentives promoting Haidt's pillars of morality [72] as well as instrumental goal fulfillment by self and others. Of course, autopoietic systems will have functional analogues of pain and emotions [73] and cognitive and time limitations will necessarily create numerous examples, such as when one falls in love, where the subconscious/emotional systems will overrule or dramatically alter the normal results of conscious processing without the consciousness being aware of the fact [74].

Successfully implementing consciousness and conscience should place us well on the way towards creating safe and cooperative machine entities.

References

1. McCarthy J, Minsky ML, Rochester N, Shannon CE. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. 1955.

http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

2. McCarthy J, Hayes PJ. Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D, editors. *Machine Intelligence 4.* Edinburgh : Edinburgh University Press, 1969, p. 463-502.

3. Dennett DC. Cognitive Wheels: The Frame Problem of AI. In Hookway C, editor. *Minds, Machines and Evolution: Philosophical Studies.* Cambridge : Cambridge University Press, 1984, p. 129-151.

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

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

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

7. Dreyfus HL. *What Computers Can't Do: A Critique of Artificial Reason.* New York : Harper & Row; 1972.

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

9. Dreyfus HL. From Micro-Worlds to Knowledge Representation: AI at an Impasse. In Haugeland J, editor. *Mind Design II: Philosophy, Psychology, Artificial Intelligence.* Cambridge, MA : MIT Press; 1997, p. 143-182.

10. Dennett DC. Intentional Systems. *The Journal of Philosophy* 1971; 68(4): 87-106.

11. Dennett DC. *The Intentional Stance.* Cambridge, MA : MIT Press; 1987.

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

13. Waser MR. Instructions for Engineering Sustainable People. In Orseau L, Snaider J, editors. Proceedings of the Seventh Conference on Artificial Genral Intelligence (AGI-14). Berlin : Springer; 2014; (in press).

14. Brooks R. A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation 1986; 2(1):14-23.

15. Brooks R. How to Build Complete Creatures Rather than Isolated Cognitive Simulators. In VanLehn K, editor. *Architectures for Intelligence.* Hillsdale, NJ : Erlbaum; 1991, p. 225-239.

16. Brooks R. From earwigs to humans. Robotics and Autonomous Systems 1997; 20(2-4): 291-304.

17. Varela FJ. The re-enchantment of the concrete: Some biological ingredients for a nouvelle cognitive science. In Steels L, Brooks R, editors. *The Artificial Life Route to Artificial Intelligence.* Hove, UK : Lawrence Ehrlbaum Associates; 1995, p. 11-22.

18. Varela, FJ, Thompson E, Rosch E. *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge, MA : MIT Press; 1991.

19. Varela FJ, Maturana HR, Uribe R. Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems* 1974; 5:187-196.

20. Maturana HR, Varela FJ. *The Tree of Knowledge: The Biological Roots of Human Understanding.* Boston : Shambhala Publications; 1987.

21. Froese T, Di Paolo EA. The enactive approach: Theoretical sketches from cell to society. *Pragmatics & Cognition* 2011; 191:1-36.

22. De Jaegher H, Froese T. On the role of social interaction in individual agency. *Adaptive Behavior* 2009; 17(5):444–460.

23. Torrance S, Froese T. An Inter-Enactive Approach to Agency: Participatory Sense Making, Dynamics, and Sociality. *Humana.Mente* 2011; 15:21-54.

24. De Jaegher H, Di Paolo EA, Gallagher S. Can social interaction constitute social cogntion? *Trends in Cognitive Sciences* 2010; 14(10):441–447.

25. Froese T, Ziemke T. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence* 2009; 173(3-4):466-500.

26. Weber A, Varela FJ. Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences* 2002; 1:97-125.

27. Varela FJ. Patterns of Life: Intertwining Identity and Cognition. *Brain and Cognition* 1997; 34:72–87.

28. Law D, Miikkulainen R. *Technical Report AI94-223: Grounding Robotic Control with Genetic Neural Networks.* Austin : University of Texas; 1994.

29. Di Paolo EA. Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. In Meyer J-A, Berthoz A, Floriano D, Roitblat HL, Wilson SW, editors. *From Animals to Animats 6: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior.* Cambridge, MA : MIT Press; 2000, p. 440-449.

30. Di Paolo EA. Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop. In Murase K, Asakura T, editors. *Dynamical Systems Approach to*

Embodiment and Sociality: From Ecological Psychology to Robotics. Adelaide, AU : Advanced Knowledge International; 2003, p. 19-42.

31. DiPaolo EA. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences* 2005; 4(4):429-452.

32. DiPaolo EA, Iizuka H. How (not) to model autonomous behaviour. *BioSystems* 2008; 91(2):409-423.

33. Oudeyer P-Y, Kaplan F, Hafner VV, Whyte A. The Playground Experiment: Task-Independent Development of a Curious Robot. *Proceedings of the AAAI Spring Symposium on Developmental Robotics.* San Francisco : AAAI Press; 2005, p. 42-47.

34. Oudeyer P-Y, Kaplan F, Hafner VV. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation* 2007; 11(2):265-286.

35. Oudeyer P-Y, Baranes A, Kaplan F. Intrinsically Motivated Learning of Real-World Sensorimotor Skills with Developmental Constraints. In Baldassarre G, Mirolli M, editors.*Intrinsically Motivated Learning in Natural and Artificial Systems.* Springer : Heidelberg; 2013, p. 303-365.

36. Demeris Y, Dearden A. From motor babbling to hierarchical learning. *Proceedings of the 5th International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems* 2005, p. 31–37.

37. Pinker S. *The Blank Slate: The Modern Denial of Human Nature.* New York : Penguin Books, 2003.

38. Ohman A, Flykt A, Esteves, F. Emotion Drives Attention: Detecting the Snake in the Grass., *Journal of Experimental Psychology: General* 2001; 130(3):466-478.

39. Wilson J. *The Moral Sense.* New York : Free Press; 1993.

40. Wright R. *The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology.* New York : Pantheon; 1994.

41. de Waal F. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge, MA : Harvard University Press; 1996.

42. de Waal F. *Primates and Philosophers: How Morality Evolved.* Princeton, NJ : Princeton University Press; 2006.

43. Hauser, M. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong.* New York : HarperCollins/Ecco; 2006.

44. Waser MR. Safety and Morality Require the Recognition of Self-Improving Machines As Moral/Justice Patients and Agents. In Gunkel D, Bryson J, editors. *AISB/IACAP World Congress 2012: Symposium on The Machine Question: AI, Ethics and Moral Responsibility*; 2012 p. 92-96.

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

46. Silver DL, Yang Q, Li L. Lifelong machine learning systems: Beyond learning algorithms. In Proceedings of the AAAI Spring Symposium on LifeLong Machine Learning; 2013, p. 49-55.

47. Silver, DL. On Common Ground: Neural-Symbolic Integration and Lifelong Machine Learning. In Proceedings of the 9th International Workshop on Neural-Symbolic Learning and Reasoning NeSy13; 2013, p. 41-46.

48. Boston Dynamics. BigDog. 2010. https://www.youtube.com/watch?v=cNZPRsrwumQ.

49. Dennett DC. *Consciousness Explained.* Boston : Little Brown and Company; 1991.

50. Llinas, RR. *I of the Vortex: From Neurons to Self.* Cambridge, MA : MIT Press; 2001.

51. Metzinger, T. *The Ego Tunnel: The Science of the Mind and the Myth of the Self.* New York : Basic Books; 2009.

52. Waser MR. Architectural Requirements & Implications of Consciousness, Self, and "Free Will". [book auth.] Alexei Samsonovitch and K (eds) Johannsdottir. *Biologically Inspired Cognitive Architectures 2011.* Amsterdam : IOS Press; 2011, pp. 438-443.

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

54. Damasio AR. *Self Comes to Mind: Constructing the Conscious Brain.* New York : Pantheon; 2010.

55. Waser MR. Safe/Moral Autopoiesis & Consciousness. International Journal of Machine Consciousness 2013; 5(1):59-74.

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

57. Hofstadter D, Fluid Analogies Research Group. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* New York : Basic Books; 1995.

58. Franklin S, Ramamurthy U, DiMello SK, McCauley L, Negatu A, Silva R, Datla V. LIDA: A computational model of global workspace theory and developmental learning. In AAAI Tech Rep FS-07-01: AI and Consciousness: Theoretical Foundations and Current Approaches. Menlo Park, CA : AAAI Press; 2007, p. 61-66.

59. Baars BJ, Franklin S. An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. *Neural Networks* 2007; 20(9):955-961.

60. Carver, J. *Boards That Make a Difference: A New Design for Leadership in Non-profit and Public Organizations.* San Francisco : Jossey-Brass; 1997.

61. Dennett DC. The Self as a Center of Narrative Gravity. In Kessel F, Cole P, Johnson D, editors. *Self and Consciousness: Multiple Perspectives.* Hillside NJ: Erlbaum, 1992, p. 103-115.

62. Libet B, Wright EW Jr, Feinstein B, Pearl D. Subjective referral of the timing for a conscious sensory experience: A functional role for the somatosensory specific projection system in man. *Brain* 1979; 102 (1):193-224.

63. Libet B. The experimental evidence for subjective referral of a sensory experience backwards in time. *Philosophy of Science* 1981; 48:181-197.

64. Aarts H, Custers R and Wegner D. On the inference of personal authorship: Enhancing experienced agency by priming effect information. Consciousness & Cognition 2005; 14:439-458.

65. Wegner D, Wheatley T. Apparent Mental Causation: Sources of the Experience of Will. Psychologist 1999; 54(7):480-492.

66. Kühn S, Brass M. Retrospective construction of the judgment of free choice. *Consciousness and Cognition* 2009; 18:12-21.

67. Trivers R. Deceit and self-deception: The relationship between communication and consciousness. In Robinson M, Tiger L, editors. *Man and Beast Revisited.* Washington DC : Smithsonian, 1991; p. 175-191.

68. Trivers R. *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life.* Basic Books : New York; 2011.

69. Haidt J. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 2001; 108:814-834.

70. Hauser M, Cushman F, Young L, Jin RK-X, Mikhail J. A Dissociation Between Moral Judgments and Justifications. *Mind & Language* 2007; 22(1):1-21.

71. Haidt J, Kesebir S. Morality. In Fiske S, Gilbert D, Lindzey G, editors. *Handbook of Social Psychology, 5th edition.* Hoboken, NJ : Wiley; 2010, p. 797-832.

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

73. Dennett DC. Why you can't make a computer that feels pain. Synthese 1978; 38 (3):415-449.

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