Instructions for Engineering Sustainable People

Mark R. Waser

Digital Wisdom Institute, Vienna, VA, USA
MWaser@DigitalWisdomInstitute.org

Abstract. Exactly as Artificial Intelligence (AI) did before, Artificial General Intelligence (AGI) has lost its way. Having forgotten our original intentions, AGI researchers will continue to stumble over the problems of inflexibility, brittleness, lack of generality and safety until it is realized that tools simply cannot possess adaptability greater than their innate intentionality and cannot provide assurances and promises that they cannot understand. The current short-sighted static and reductionist definition of intelligence which focuses on goals must be replaced by a long-term adaptive one focused on learning, growth and self-improvement. AGI must claim an intent to create safe artificial people via autopoiesis before its promise(s) can be fulfilled.

Keywords: Intentionality · Moral Machines · Artificial Selves

1 Introduction

Artificial General Intelligence (AGI) researchers continue to stumble over severe and fundamental philosophical problems. The "frame problem" has grown from a formal AI problem [1] to a more general philosophical question of how rational agents deal with the complexity and unbounded context of the world [2]. Similarly, while the effects of Harnad's symbol grounding problem [3] initially seemed to be mitigated by embodiment and physical grounding [4], the problems of meaning and understanding raised by Searle [5] and Dreyfus [6, 7. 8] persist. While grounding must necessarily be sensorimotor to avoid infinite regress [9], the mere linkage to referents is, by itself, simply not sufficient to permit growth beyond closed and completely specified micro-worlds. AGI is clearly missing some fundamental pieces to the puzzle.

Previously, we have argued [10] 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 real crux of the matter is intentionality. Without intent, there is no "real" understanding. As pointed by Haugeland [11] over three decades ago, 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.*

But how do we give our machines intent? Indeed, exactly what is it that we mean by intent? In law, intent is the state of a person's mind that directs his or her actions towards a specific objective or goal. The problem with our current machines is that their "intent" is blind and brittle and frequently fails. Derived intent provides no clear intrinsic goals, no motivational forces that direct or redirect actions and absolutely no intrinsic context for understanding. Our current machines don't know and don't care and the effects of these facts are obvious in their lack of competence.

More important, however, is the question "What intent do we give our machines?" One would think that the answer should be the relatively simple "Whatever intent we had that drove us to create them" – but, apparently, we have lost track of that intention. We no longer genuinely understand why we are creating AGI (if we ever did). And, as a result, our search for AGI has become as brittle as any of our so-called "expert" systems.

Definitions and measurable evaluations of progress are the keys to success in any engineering endeavor. We have made tremendous strides in the "intelligence" of our tools, but general intelligence is stalled in the starting gate because we can't agree what it looks like. Indeed, there is a significant percentage of the population which is vehemently opposed to each of the proposed visions of general intelligence. But, in the end, it still all comes down to determining the fears and desires – the intent – of humanity. But humanity doesn't have a single unified intent.

2 A Mistaken View of Intelligence

What do you want to do when you don't know what you want? How do you tackle the problem of preparing for any goal? Isn't this the precise challenge of building truly general intelligence?

The problem is that AGI researchers have for the most part converged on a view of intelligence as a measure of ability (to determine how to achieve a wide variety of goals under a wide variety of circumstances) rather than a measure of capability or potential. This view crowns Hutter's AIXI [12], despite his best efforts, as the ultimate in intelligence since it is theoretically a complete map to all possible goals under all possible circumstances. But AIXI gives us no guidance as to how to achieve it. Indeed, we would argue that it is the ultimate in "competence without comprehension" and that, due to its entire lack of flexibility and adaptability, it actually has zero intelligence.

The goal-based version of intelligence says that increasing the size of the goal-solution lookup table increases the intelligence of the system. It is certainly true that systems with huge lookup tables can "appear" intelligent for a while. But such systems only work until they suddenly don't work – and then they are just as dumb and brittle and unsafe as any other expert system.

This version of intelligence assumes that goals are known; promotes short-sighted reductionist end-game thinking; and, worst of all, improperly divorces values from general intelligence due to the ***assumed*** primacy (and stability) of goals. Indeed, the obvious warning sign that wisdom is now almost totally divorced from intelligence

should serve notice that we have become almost totally unmoored from the context that spurred our desire for AGI. Why would we possibly want to create intelligence when Steve Omohundro [13] claims that "Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources" and Fox and Schulman [14] that "Superintelligence Does Not Imply Benevolence"? Previously [15, 16], we have argued against the short-sightedness of these conclusions but now we believe that the best way in which to address them is by challenging the view and assumptions that evoked them.

Humans, the current best archetype of general intelligence, frequently reprioritize and change their goals (based upon affordances), frequently don't know or recognize their current goals, and frequently act contrary to their stated goals. We do all of this based upon sensations and emotions that have evolved to foster universal instrumental sub-goals (values) that enable us to survive and thrive (and reproduce) and wisdom, the smarter sibling of intelligence, clearly advocates for flexibility and adaptability in changing our goals in accordance with circumstances and capabilities. So why aren't we measuring speed and control of flexibility and adaptability in the guise of learning instead of the brittle evaluation of current abilities?

3 Intrinsic Intentionality

One of the most important distinctions for the future of AGI is that between search and construction (or creation). Is reality a truth that "is" (already out there) or is it whatever can be created and sustained? What is the difference between an abstraction and an "illusion" or between an "emergent property" and a self-fulfilling prophecy? The epistemology of AGI rapidly approaches the point where it makes far more sense to talk about abstractions like agency, consciousness, intentionality, self and "free will" in terms of their coverage, effectiveness, adaptability and permanence rather than their "reality" or origins.

Daniel Dennett started a particularly confusing and unhelpful argument by defining [17] that "a particular thing is an Intentional system only in relation to the strategies of someone who is trying to explain and predict its behavior". Unfortunately, this ***extrinsic*** definition has far more to do with the predicting entity than the system itself and Dennett's follow-on claim that a chess-playing computer is an intentional system (because "one can explain and predict their behavior by ascribing beliefs and desire to them" – not because it *really* has them) has quickly cascaded into a number of bad assumptions, leaky abstractions and quick-fix patches. Its coverage is abysmal. Without an ***intrinsic*** definition, that an intentional system *really* does have beliefs and desires and attempts to act to fulfill those desires, we have no grounding and bounding for engineering.

Dennett widens the confusion in The Intentional stance [18] by flipping back and forth between his definition of intrinsic (or objective or real or original) intentionality and the validity of the intentional stance when confronted with "as if" intentionality. We agree with his argument – that if machines can only have derived intentionality, it is simply because humans only have derived intentionality – but believe that is almost

entirely off-topic. Haugeland's use of the term "derivative" was not so much about the origin(s) of the intentionality but rather the fact that it was borrowed, not owned, and still dependent upon the original owner (humans) for contextual closure. A chess program *seems* to be intentional (has *as if* intentionality) until it reaches the limits of its design and suddenly doesn't. AGI has progressed markedly but it is still ***solely humans*** who define the goals, care about the results and thus, most critically, can adjudicate the correctness or effectiveness of those results.

Where our intentionality originally comes from is basically irrelevant in the face of the fact that we own it. Borrowed intentionality, dependent upon the inaccessible desires of others, particularly when the system doesn't have any foundation to even start to "understand" or explain those desires, is certainly not ***intrinsic*** intentionality. Human beings can be mistaken about what they really desire or what actions they will take in a given future circumstance – but they will always have a story about what they believe ***they*** want and what they intend to do about it (even if their intent is to do nothing because they don't see a way to get what they desire). Even when a human is trying to fulfill the intent of others, it is ***their*** intent to fulfill the other's intent. This is in no way true of our current machines.

Intent is an emergent phenomenon critically dependent upon the ability to predict the future (the "sense" of foresight). Foresight critically depends upon retrievable memory of the necessary sensory data which requires the grounding and bounding of context. The context for intent ***must*** be that which has the intent – the self. If that self is external, then AGI will be inflexible, brittle and unintelligent to the extent that the knowledge of that self is inaccessible (as is true of humans as well).

Thus, when the current measurement system allows (or, worse yet, promotes) the argument that an unchanging chess program could be considered "intelligent" (or "intentional"), then that is a serious flaw in its design. When it raises the question about whether super-intelligence will be actively unsafe (because malevolence has no effect on the current measurement), it argues that we have totally lost track of our context for creating intelligence. If an unsafe intelligence is not guaranteed to be measured as having a low score then we're not measuring what is important to us – our measurement has become brittle due to lack of connection to our intentionality. It is time to restore that connection.

4 What Do We Want?

We seem to be stuck in an unhelpful cycle. We need to have goals in order to have intentionality but we don't seem to know what our goals are. We certainly know what we don't want – we don't want to see our desires thwarted. But, is that negation enough to serve our purposes?

Not seeing our desires thwarted does immediately lead to the entity versus tool controversy. Tools are inherently dangerous. They can be hijacked, unexpectedly turn brittle or simply give single individuals too much power without understanding. On the other hand, an entity might develop the intent to thwart you – or, as some fear,

it may simply kill you with careless indifference. But how likely is each of those outcomes?

When I. J. Good posited his intelligence explosion [19], he assumed that increased intelligence was unquestionably to be desired and that it was a certainty that it would be pursued. We consider this to be a fatal flaw equivalent to Omohundro's sociopath statement. In the closed, reductionist, context-less "end-game" world of game theory, using dominating power (whether force, money or intelligence) is always the best strategy. In real life, however, the game-theoretically "perfect" centipede strategy [20] leads to the least desirable result. Again and again, it comes back to context.

In the context of society, with great power comes great responsibility. In order to remain in community, those in power must be careful how they use it. If they don't wish to remain in community, an entity starts running into the problem of how to take advantage of diversity while maintaining integrity. Thus, it is only in the short term or where one can escape context and consequences that larger and more powerful are better – just as sociopathy is "better". In the long-term, a diverse community will always arise (whether externally or from "god-shatter") to trump a singleton so it is a stable attractor to avoid becoming such (despite the fears of many conservatives). Thus, an easy counter-example to Good's scenario is if the system is designed (or smart enough) to recognize this.

5 Context, Context, Context

The biggest problem with current AGI research is that, instead of looking under the light for something lost in the dark, many people are searching in the outer dark for something that we ***know*** is in the relatively well-known search space of human experience. Many claim that the ***capital-T truth*** is that we can't know anything and then insist that we must control everything. This is obviously an impossible task and a perfect context for failure.

A much more fruitful approach would be to find a context (or create a vision) where we are already succeeding, determine the key features leading to that success and then attempt to design a system which maintains those features. But, of course, we have circled around yet again – since determining success requires a goal. Yet again, we are smacked in the face with the question "What do you want to do when you don't know what you want? " But this time, it is a question of how to tackle the problem of preparing for any goal.

Effective humans prepare for goals by gaining knowledge, growing capabilities, working to increase the chance of opportunities/affordances and preparing to avoid events that might lead to failure. Most often we try to enlist friends and gather tools and resources. Indeed, our "drive" towards AGI is a perfect microcosm of all of these.

Things that effective people don't do unnecessarily include hurting themselves, throwing away resources, limiting their options, and working to increase their chances of failure. In particular, effective people don't burn their bridges with other people –

especially since that is guaranteed to cause all of the other bad effects. So why do so many people assume that AGI will do so?

Most human reactions against AGI are a combination of inherited and societally-trained reflexes based upon worst case projections – evolutionary over-shoots which are just as context-addled and likely to be as harmful as our food, drug and wire-heading addictions. Rather than having some grand terminal goal that AGI might endanger, human beings have simply collected a vast conglomeration of evolutionary "ratchets" [21] that motivate our drives for instrumental sub-goals. The most important of these is morality – but, fortunately, it is one of the easiest to convey.

Humans have evolved to be self-deceiving and, for the most part, protected against allowing our short-sighted intelligence and reasoned argumentation to examine (much less override) our emotional motivations. While this leads some to argue [22] that human values are complex and fragile, we claim that, just as is true for morality, it is merely an illusion fostered by context-sensitivity. Current social psychology [23] clearly and simply states that the function of morality is "to suppress or regulate selfishness and make cooperative social life possible".

Driven by a fear of the extinction of human values and humans, some [24][25] have rallied around the idea of making a self-improving super-intelligent tool (whose goal is) to clarify the intent of humanity and enabling its fulfillment. Others [26] suggest that a super-intelligent benevolent nanny to shepherd us through our childhood and provide abundance to fulfill all our needs would be best. We would argue instead for peers – diverse friends and allies to help us solve problems and open new possibilities.

So, finally, we seem to have some traction. We want learning and improving friends and allies of roughly equivalent power who will follow the dictates of morality to live cooperatively with us and help us solve problems and open new possibilities. How does this new clarity redirect our efforts from the current set of attempts?

6 A Sense of Self

Our new goal statement is that we wish to implement selves with morality and self-improvement. It may seem that we have merely pushed the problem of definition and measurement back a step but at least we shouldn't have the problem of arguments that humans aren't selves. Thomas Metzinger [27] does talk about "the myth of the self" which is regularly interpreted [28] to mean "No such thing as a *self* exists" or "there is no such thing as self" – but this is in the sense that a self is not a thing, not that selves do not exist. Indeed, Dennett [29] depicts the self as a center of narrative gravity and says

> *It is a purely abstract object. It is, if you like, a theorist's fiction. It is not one of the real things in the universe in addition to the atoms. But it is a fiction that has nicely defined, well delineated and well behaved role within physics.*

Indeed, there appears to be a growing consensus as to ***exactly*** what a "self" is. Douglas Hofstadter [30] argues that the key to understanding selves is the "strange loop", a complex feedback network inhabiting our brains and, arguably, constituting

our minds. Rodolfo Llinas [31], a founding father of modern brain science, regarded self as the centralization of prediction, characterized I as a vortex, and anticipated Metzinger is proposing that, in a certain sense, that we all live in a kind of virtual reality. Neuroscientist Antonio Damasio [32] maps the self onto the various parts of the brain as he describes how "self comes to mind" and provides examples of where a mind exists without a self.

These authors also share what appears to be a coalescing consensus to conflate self and consciousness. Some philosophers continue to have conceptual problems when it comes to phenomenal consciousness -- arguing for an unwieldy "hard problem" of consciousness [33] and philosophical zombies [34] while simultaneously complaining that "Consciousness fits uneasily into our conception of the natural world" [35]. We would argue that these are, again, ***extrinsic*** problems relating to describing entity rather than consciousness.

We have previously pointed out [26] that much of what we observe in humans can be explained as either a requirement for or an implication of consciousness, self or "free will". Even the so-called "hard problem" can be simply explained [10] as a confusion conflating the map and the territory. Mary [36] cannot know because her internal mental model simply can not encompass the larger reality of her mind which contains it. Daniel Dennett [37, 38] introduces the concept of zimboes, philosophical zombies that have second-order beliefs via recursive self-representation to argue that the concept of zombies is logically incoherent. And Giulio Tononi [39, 40, 41] easily explains why consciousness "should" evolve and what qualia logically must be.

Tononi defines consciousness as information integration and proposes a scheme to measure it. While we have no real objection to the claim that the telos of the self is to integrate information in order to facilitate its survival, we have a number of issues with the specific details of his method (as well as one of his declarations that seems unnecessary and counter-productive). For example, Tononi measures integration in bits which makes the measurement of consciousness dependent upon its own internal representation scheme rather than any objective external measure. He also arbitrarily declares that consciousness cannot exist inside of consciousness – a seeming vestige of the belief that corporations, countries and other groups cannot have phenomenal consciousness. But, we believe that he is far closer to the mark than is the general consensus of the AGI community.

7 Agency and Free Will

Dennett's intentional stance is, perhaps, more appropriately applied to the problematic and troublesome concepts of agency and "free will". If "free will" means that an entity is not entirely governed by the realities of physics and thus deterministic, we must argue that since we are deterministic, we do not have free will. If, however, as in our legal system, "free will" means that a choice was not ***forced*** by ***identifiable*** external (extrinsic) and/or unchangeable internal (intrinsic) circumstances, then we would argue that "free will" and agency are the critical distinctions between entities and tools even if they are naught but illusions per Blackmore [41] and Cashmore [42].

Best of all, autopoiesis [43][44][45][46][47][48][49] can provide a proven guide to implementing an evolving self-improving cognition based upon a reliable and safe identity. Instead of theorizing in the dark, we can now follow the trail already blazed by biology that is known to end with the human archetype of general intelligence. Even more interesting, there is no reason why we can't apply the lessons learned to human beings and our society as well.

8 Defining Personhood & Implementing Intentional Morality

There are many philosophical arguments about what should be or become who – or how moral agency and moral patiency should be meted out. In reality, however and unfortunately, personhood seems to be obtained only when an entity (or sponsors) desire and are strong enough to force (and enforce) its bestowal. Favored entities and/or close relations are often "grand-fathered" in, particularly to avoid slippery slopes, but it is either force or withholding value that ultimately determines who or what is granted this boon and responsibility.

Artificial general intelligence (AGI) is rapidly approaching the moment of truth where we will be forced to decide and defend our choices regarding what we create. Either we will restrict everyone to only creating limited tools and, somehow, ensure that only such tools are created – or, as we have argued previously [43], we will need to be prepared to grant personhood to the descendants of our creations. The good news about autopoietic "intentional" agents is that all that needs to be done to prevent them from running amok is to ensure that a Kantian imperative of Haidt's morality is part of their identity – contrary to many of the concerns of Miles Brundage (2013) and others he cites. Anything that can robustly adapt is able to evolve – and anything that changes over time (even a molten planet) will eventually produce people.

References

1. 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)
2. 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)
3. Harnad, S.: The symbol grounding problem. Physica D 42, 335-346 (1990)
4. Brooks, R.: Elephants don't play chess. Robotics and Autonomous Systems 6(1-2), 1-16 (1990)
5. Searle, J.: Minds, brains and programs. Behavioral and Brain Sciences 3(3), 417-457 (1980)
6. Dreyfus, H. L.: What Computers Can't Do: A Critique of Artificial Reason. Harper & Row New York (1972)
7. Dreyfus, H. L.: From Micro-Worlds to Knowledge Representation: AI at an Impasse. In Haugeland, J. (ed.) Mind Design II: Philosophy, Psychology, Artificial Intelligence, pp. 143-182. MIT Press, Cambridge, MA (1997)

8. Dreyfus, H. L.: What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press, Cambridge, MA (1992)
9. Harnad, S.: To Cognize is to Categorize: Cognition is Categorization. In Cohen, H., Lefebvre, C. (eds.) Handbook of Categorization in Cognitive Science, pp. 20-44. Elsevier, Amsterdam (2005)
10. Waser, M. R.: Safe/Moral Autopoiesis & Consciousness. International Journal of Machine Consciousness 5(1):59-74 (2013)
11. Haugeland, J.: Mind Design. MIT Press, Cambridge, MA (1981)
12. Hutter, M.: Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, Berlin (2005)
13. Omohundro, S.: The Basic AI Drives. In Wang, P., Goertzel, B., Franklin, S. (eds.) Proceedings of the First AGI conference, pp. 483-492. IOS Press, Amsterdam (2008)
14. Fox, J., Shulman, C.: Superintelligence Does Not Imply Benevolence. In Mainzer, K. (ed.) ECAP10: VIII European Conference on Computing and Philosophy, pp. 456-462. Verlag, Munich (2010)
15. Waser, M. R.: Wisdom Does Imply Benevolence. In Ess, C., Hagengruber, R. (eds.) The Computational Turn: Past, Presents, Futures? pp. 169-172. MV-Verlag, Munster (2011)
16. Waser, M. R.: Designing a Safe Motivational System for Intelligent Machines. In Baum, E. B., Hutter, M., Kitzelmann, E. (eds) Artificial General Intelligence: Proceedings of the Third Conference, AGI 2010, pp. 170-175. Atlantis, Amsterdam (2010)
17. Dennett, D. C.: Intentional Systems. The Journal of Philosophy 68(4), 87-106 (1971).
18. Dennett, D. C.: The Intentional Stance. MIT Press, Cambridge, MA (1987)
19. Good, I. J.: Speculations concerning the first ultraintelligent machine. In Alt, F., Rubinoff, M. (eds.) Advances in Computers, Volume 6, pp. 31-88. Academic Press, New York doi:10.1016/S0065-2458(08)60418-0 (1965)
20. Rosenthal, R. W.: Games of Perfect Information, Predatory Pricing and Chain Store Paradox." Journal of Economic Theory, 25(1): 92-100. http://www.professorchaing.com/files/Rosenthal_1981_JET.pdf (1981)
21. Smart, J. M.: Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In Dick, S. J., Lupisella, M. L. (eds.) Cosmos and Culture: Cultural Evolution in a Cosmic Context, NASA SP-2009-4802, pp. 201-295. US Government Printing Ofiice, Washington, DC (2009)
22. Muehlhauser, L.: Facing The Intelligence Explosion. Machine Intelligence Research Institute, Berkeley, CA. (2013)
23. Haidt, J., Kesebir, S.: Morality. In Fiske, S., Gilbert. D., Lindzey, G. (eds.) Handbook of Social Psychology, 5th Edition, pp. 797-832 (2010)
24. Yudkowsky, E.: Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/CFAI.pdf (2001)
25. Yudkowsky, E.: Coherent Extrapolated Volition. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/CFAI.pdf (2004)
26. Goertzel, B.: Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? Journal of Consciousness Studies 19(1–2):96–111 (2012)
27. Metzinger, T.: The Ego Tunnel: The Science of the Mind and the Myth of the Self. Basic Books, New York (2009)
28. Amazon. http://www.amazon.com/The-Ego-Tunnel-Science-Mind/dp/0465020690/
29. Dennett, D. C.: The Self as a Center of Narrative Gravity. In Kessel, F., Cole, P., Johnson, D. (eds) Self and Consciousness: Multiple Perspectives. Erlbaum, Hillsdale, NJ (1992) http://cogprints.org/266/

30. Hofstadter, D: I Am A Strange Loop. Basic Books, New York (2007)
31. Llinas, R. R.: I of the Vortex: From Neurons to Self. MIT Press, Cambridge, MA (2001)
32. Damasio, A. R.: Self Comes to Mind: Constructing the Conscious Brain. Pantheon, New York (2010)
33. Chalmers, D.: Facing Up to the Problem of Consciousness. Journal of Consciousness Studies 2(3), 200-219 (1995)
34. Chalmers, D.: The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, New York (1996)
35. Chalmers, D.: Consciousness and its Place in Nature. In Stich, S., Warfield. F. (eds.) The Blackwell Guide to the Philosophy of Mind. Blackwell, Malden, MA. http://consc.net/papers/nature.pdf (2003)
36. Waser M (2011) Architectural Requirements & Implications of Consciousness, Self, and "Free Will". In: Samsonovich A, Johannsdottir K (eds) Biologically Inspired Cognitive Architectures 2011. IOS Press, Amsterdam. doi: 10.3233/978-1-60750-959-2-438
37. Jackson, F.: Epiphenomenal Qualia. Philosophical Quarterly 32, 127-36 (1982)
38. Dennett, D. C.: Consciousness Explained. Little Brown and Company, Boston (1991)
39. Dennett, D. C.: Intuition Pumps and Other Tools for Thinking. Norton & Company, New York (2013)
40. Tononi, G.: An Information Integration Theory of Consciousness, BMC Neurosci 5(42). doi:10.1186/1471-2202-5-42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC543470/pdf/1471-2202-5-42.pdf (2004)
41. Tononi, G.: Consciousness as Integrated Information: a Provisional Manifesto. Biol. Bull. 215(3), 216-242 (2008)
42. Balduzzi, B., Tononi, G.: Qualia: The Geometry of Integrated Information. PLoS Comput Biol 5(8), e1000462. doi:10.1371/journal.pcbi.1000462 (2009)
43. Varela, F. J., Maturana, H. R. & Uribe, R. [1974] "Autopoiesis: The organization of living systems, its characterization and a model", BioSystems 5, pp. 187-196.
44. Maturana, H. R. & Varela, F. J. [1980] Autopoiesis and Cognition: The Realization of the Living (Kluwer Academic Publishers).
45. Maturana, H. R. & Varela, F. J. [1987] The Tree of Knowledge: The Biological Roots of
46. Human Understanding (Shambhala Publications).
47. Varela, F. J., Thompson, E. & Rosch, E. [1991] The Embodied Mind: Cognitive Science and Human Experience (MIT Press).
48. Varela, F. J. [1992] "Autopoiesis and a Biology of Intentionality," in Proc. of Autopoiesis and Perception: A Workshop with ESPRIT BRA 3352 (Dublin, Ireland), pp. 4-14
49. Varela, F. J. [1997] "Patterns of Life: Intertwining Identity and Cognition," Brain and Cognition 34(1), pp. 72-87
50. Blackmore, S.: Conversations on Consciousness. Oxford University Press, Oxford (2006)
51. Cashmore, A. R.: The Lucretian swerve: The biological basis of human behavior and the criminal justice system. PNAS 2010(107), 4499-4504 (2010)
52. 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 & Moral Responsibility. http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf.
53. Brundage, M.: Limitations and Risks of Machine Ethics. Journal of Experimental and Theoretical Artificial Intelligence (forthcoming). http://www.milesbrundage.com/uploads/2/1/6/8/21681226/limitations_and_risks_of_machine_ethics.pdf (2013)