A Myriad of Automation Serving a Unified Reflective Safe/Moral Will

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

Abstract

We propose a unified closed identity with a pyramid-shaped hierarchy of representation schemes rising from a myriad of tight world mappings through a layer with a relatively small set of properly integrated data structures and algorithms to a single safe/moral command-and-control representation of goals, values and priorities.

Rephrasing the Question

The question "How should intelligence be abstracted", when followed by a list of representation schemes, implies and thereby continues to promote a trio of questionable assumptions that we contend have hampered the creation of a truly general artificial intelligence for years. First, it focuses on the specific form of the static product of abstraction, the representation scheme, as if it were the missing key to intelligence, rather than the process of abstraction itself. Second, such a focus continues the emphasis on the analysis and creation of "intelligent" tools (which we consider an oxymoron) rather than creating a complete (and general) intelligence -- a unified identity (or self) satisfying a closure property. Finally, it implies that the minimal core of intelligence is best implemented using but a single scheme or a small serial set rather than using a larger number of representations, particularly in parallel.

We believe that the key to intelligence is not in the details of representation but in implementing the capability of abstracting from one level or representation scheme to the next – particularly if the higher level can utilize multiple lower levels in parallel or even cooperatively. Indeed, we would argue that, to be most effective, intelligence should possess a pyramid-shaped hierarchy of representation schemes with the lowest levels using many disparate schemes to map as closely as possible to the features of the world, the middle levels comprising a toolbox or "cognitive substrate" of parallel cooperative processes and schemes, and the peak being a singular command-and-control representation scheme of goals, values and priorities. Instead of asking "How should intelligence be abstracted", the question should be "How can different abstractions be brought together cooperatively to create a true general intelligence?"

A Biologically Inspired Architecture

How is it that a human being can approach, understand and solve virtually any problem? Many point at the human brain and claim that neurons and neural networks are the obvious answer as the archetypal example of a usable "one-size-fits-all" representation scheme. However, this is as helpful as the equally accurate statement that ones and zeros (or machine code) could do so as well. An incredible diversity of neurons can (and do) combine in a myriad of ways to implement virtually any higher level abstraction or representation scheme desired – but this diversity also rules out coherently considering them a single uniform scheme.

Instead, the human mind is most effectively regarded as two separate systems: the slow, serial, symbolic, pseudo-logical, reflective consciousness and the fast, parallel, sub-symbolic, opaque, constraint-satisfaction subconscious (Kahneman 2011). Thus, we would argue for a representational architecture based upon that distinction (Baars and Franklin 2007) as we have discussed previously (Waser 2012).

The simplest minds are reflexive, merely reacting in response to immediate stimuli. Slightly more advanced minds contain world models so that expected problems and the unexpected become "immediate stimuli" that can be reacted to. Still more advanced minds show the earliest signs of learning by altering their reactions if their results are persistently negative. Continuing this trend, towards circularity and necessary reflection and its implications (Waser 2011), leads to Richard Dawkins' (1976) speculation that "perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself."

Thus it is not surprising that Hofstadter (2007) argues that the key to understanding consciousness and (our) selves is the "strange loop", the complex feedback network inhabiting our brains and, arguably, constituting our minds. Indeed, we would argue that consciousness is intelligence and that everything else is but reflex and automation (tools) created either initially by evolution or, in conscious minds, by automatization (Franklin et al 2007).

The Shortcomings of Symbolic AI

The term "Good Old-Fashioned AI (GOFAI)" was coined to declare that symbol manipulation alone was insufficient to create intelligence capable of dealing with the real world. The "frame problem" (McCarthy & Hayes 1969, Dennett 1984), Searle's (1980) "Chinese Room", and Harnad's (1990) "symbol grounding problem" seemingly prevent GOFAI from growing beyond closed and completely specified micro-worlds. While some fully grounded and bounded systems have had spectacular successes in endeavors ranging from beating chess grand masters to autonomous driving (of course, only to subsequently be declared not to be "true AI" – correctly in our opinion as they are merely reactive compiled tool), many others have failed in equally spectacular fashion.

Part of the problem is soluble through embodiment or the use of linked sub-symbolic systems to sense and map the world and predict how it will behave (physical grounding). But this still leaves the larger part of the problem. Indeed, Perlis (2010) claims that

> Rational anomaly-handling (RAH) is then the missing ingredient, the missing link between all our fancy idiot-savant software and human-level performance. Notice the boldness of this claim: not simply do our systems lack RAH, but this lack is the missing ingredient.

However, it is our claim that RAH is impossible with the existential grounding and bounding provided by closure. As long as intentionality (Dennett 1987) is implicit and derivative from humans rather than self-contained and explicit, we will not be able to create flexible, truly general systems that we can also ensure will be friendly to humans.

The Evolved Solutions of Consciousness

Cassimatis (2006) points to evidence from linguistics, cognitive psychology, and neuroscience to claim that "a relatively small set of properly integrated data structures and algorithms can underlie the whole range of cognition required for human-level intelligence" and that "once the problems of artificial intelligence are solved for these" then "the rest of human-level intelligence can be achieved by the relatively simpler problem of adapting the cognitive substrate to solve other problems".

The evolution of attention (Ohman, Flykt, and Esteves 2001) demonstrates a solution to AI's temporal problems. Also, consciousness's symbolic nature allows us to create a grounded world model with fixed pleasure, pain, curiosity and dissonance points similar to humanity's evolved innate traits (Pinker 2003) to ensure that the machine has the same moral sense and safe motivational system that humans have (Waser 2010) rather than a dangerous system of short-sighted, un-evolved drives (Omohundro 2008).

References

Baars, B.J. and Franklin, S. 2007. An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. Neural Networks 20:955-961.

Cassimatis, N.L. 2006. A Cognitive Substrate for Achieving Human-Level Intelligence. AI Magazine 27(2): 45-56.

Dawkins, R. 1976. The Selfish Gene. Oxford: Oxford Univ Press.

Dennett, D. 1987. The Intentional Stance. Cambridge, MA: MIT Press.

Dennett, D. 1984. Cognitive Wheels: The Frame Problem of AI. In Hookway, C. ed. *Minds, Machines & Evolution: Philosophical Studies*, 129-151. New York, NY: Cambridge University Press.

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

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

Franklin, S.; Ramamurthy, U.; D'Mello, S.; McCauley, L.; Negatu, A.; Silva R.; and Datla, V. 2007. LIDA: A computational model of global workspace theory and developmental learning. In AAAI Tech Rep FS-07-01: 61-66. Menlo Park, CA: AAAI Press.

Kahneman, D. (2011) Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux.

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

Omohundro, S. 2008. The Basic AI Drives. In Wang, P.; Goertzel, B.; and Franklin, S. eds. Proceedings of the First Conference on Artificial General Intelligence. Amsterdam: IOS.

Perlis, D. 2010. BICA and Beyond: How Biology and Anomalies Together Contribute to Flexible Cognition. International Journal of Machine Consciousness 2(2): 1-11.

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

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

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

Waser, M. R. 2012. Safely Crowd-Sourcing Critical Mass for a Self-Improving Human-Level Learner/"Seed AI". In Biologically Inspired Cognitive Architectures 2012: 345-350. Berlin: Springer.

Waser, M. R. 2011. Architectural Requirements & Implications of Consciousness, Self, and "Free Will". In Biologically Inspired Cognitive Architectures 2011: 438-443. Amsterdam: IOS Press.

Waser, M. R. 2010. Designing a Safe Motivational System for Intelligent Machines. In Proceedings of the Third Conference on Artificial General Intelligence: 170-175. Amsterdam: Atlantis.