Safely Crowd-Sourcing Critical Mass for a Self-Improving Human-Level Learner/"Seed AI"

Mark R. Waser

Abstract

Artificial Intelligence (AI), the "science and engineering of intelligent machines", still has yet to create even a simple "Advice Taker" (McCarthy 1959). We argue that this is primarily because more AI researchers are focused on problem-solving or rigorous analyses of intelligence rather than creating a "self" that can "learn" to be intelligent and secondarily due to the excessive amount of time that is spent re-inventing the wheel. We propose 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 learner starting anywhere above the critical knowledge will acquire the vital knowledge that a typical human learner would be able to acquire. We believe that a ***moral***, self-improving learner ("seed AI") can be created today via a safe "sousveillance" crowd-sourcing process and propose a plan by which this can be done.

"Learning" to become intelligent

While the verb "to learn" has numerous meanings in common parlance, for the purposes of this paper, we will explicitly define a "learner" solely as a knowledge integrator. In particular, this should be considered as distinct from a "discoverer", a "memorizer", and/or an "algorithm executor" (although these are all skills that can be learned). Merely acquiring knowledge or blindly using knowledge is not sufficient to make a learner. Learning is the functional integration of knowledge and a learner must be capable of integrating all acquired knowledge into its world model and skill portfolio to a sufficient extent that it is both immediately usable and can also be built upon.

Recently, it has been hypothesized (Samsonovich 2011) that for a large set of learning environments and setting, there is one minimal set of initial cognitive and

learning characteristics (called critical mass), such that a learner starting below the critical mass will remain limited in its final knowledge by the level at which it started, while a learner starting anywhere above the critical mass will acquire the vital knowledge that a typical human learner would be able to acquire under the same settings, embedding and paradigms. Effectively, once a learner truly knows how to learn, it is capable of learning anything, subject to time and other constraints. Thus, a learner above critical mass is a "seed AI", fully capable of growing into a full blown artificial intelligence.

It has been pointed out (Waser 2011) that the vast majority of AGI researchers are far more focused on the analysis and creation of intelligence rather than self and generally pay little heed to the differences between a passive "oracle", which is frequently perceived as not possessing a self, and an active autonomous explorer, experimenter, and inventor with specific goals to accomplish. We simply point out that, in order to self-improve, there must be a self. By focusing on a learner, we can to answer or avoid many of the questions that derail many AI researchers. We will draw on the human example while remembering that many aspects and details of the human implementation of learning are clearly contra-indicated for efficiency or safety reasons and many common debates can be ignored as "red herrings" that don't need to be pursued. We are not attempting to create intelligence – whatever that is. We are creating a safe learner, a knowledge integrator that will not endanger humanity.

Objects and Processes

"Self" and "consciousness" are two primary examples of what Marvin Minsky calls "suitcase words" – words that contain a variety of meanings packed into them (Minsky 2006). For the purposes of this paper, we will consider them solely from the point of view of being functional objects and functional processes. We will handle "morality" similarly as well, using the social psychology definition that states that the function of morality is "to suppress or regulate selfishness and make cooperative social life possible" (Haidt and Kesebir 2010).

A learner is self-modifying and the complete loop of a process (or a physical entity) modifying itself must, particularly if indeterminate in behavior, necessarily and sufficiently be considered as an entity rather than an object – which humans innately tend to do with the pathetic fallacy. "I Am a Strange Loop" (Hofstadter 2007) argues that the mere fact of being self-referential causes a self, a soul, a consciousness, an "I" to arise out of mere matter but we believe that this confuses the issue by conflating the physical self with the process of consciousness. The "self" of our learner will be composed of three parts: the physical hardware, the personal memory/knowledge base, and the currently running processes.

Information integration theory claims (Tononi 2004) that consciousness is one and the same thing as a system's capacity to integrate information – thus providing both the function of consciousness and a measure. We disagree, however, with

the contentions that when considering consciousness, we should "discard all those subsets that are included in larger subsets having higher Φ (since they are merely parts of a larger whole)" and that collections of conscious entities aren't conscious either – which seem to be the results of some sort of confirmation bias that there is something "special" about human-scale consciousness. Rather than considering only simple connected graphs, we believe that the theory needs to be extended to consider modularity, encapsulation, and the fact that subsystems virtually never pass on all information. We believe that both the oft-debated questions "Is the human subconscious conscious?" and "Is the US conscious?" are answered with a clear yes merely by observing that they clearly perform information integration at their individual system level.

Arguably, there is a widely scaled range of encapsulated and modular systems that integrate information. Human minds are already often modeled as a society of agents (Minsky 1988) or as a laissez-faire economy of idiots (Baum 1996). The ***argumentative theory*** (Mercier and Sperber 2001) looks like an exact analogy one level higher with groups or society as a whole being the mind and confirmation-biased individuals merely contributing to optimal mentation. Indeed, many of the ideas proposed for our logical architecture were inspired by or drawn directly from one of the newer models in organizational governance (Carver 2006).

Critical Components

Self-Knowledge/Reflection - A "self" is not truly a self until it knows itself to be one. In this case, the self will be composed of three parts: the running processes, the personal memory/knowledge base, and the physical hardware. The learner will need to start with a competent model of each as part of its core knowledge base and sensors to detect changes and the effects of changes to each.

Explicit Goals - A learner is going to be most effective with the explicit goal to "acquire and integrate knowledge". On the other hand, the potential problems with self-modifying goal-seeking entities have been amply described (Omohundro 2008). Therefore, as per our previous arguments, in order to be safe, the learner's topmost goal **must** be the "moral" restriction "Do not defect from the community" (Waser 2012).

Self-Control, Integrity, Autonomy, Independence & Responsibility - The learner needs to be in "predictive control" of its own state and the physical objects that support it – being able to consistently predict what generally will or will not change and fairly exactly what those changes will be. This is one area where our learner will deviate markedly from the human example in a number of significant ways in order to answer both efficiency and safety concerns. Humans have evolved to self-deceive in order to better deceive others (Trivers 1991). Indeed, our evolved moral sense of sensations and reflexive emotions is almost entirely

separated from our conscious reasoning processes with scientific evidence (Hauser et al. 2007) clearly refuting the common assumptions that moral judgments are products of, based upon, or even correctly retrievable by conscious reasoning. Worse, we don't even really know when we have taken an action since we have to infer agency rather than sensing it directly (Aarts et al. 2005) and we are even prone to false illusory experiences of self-authorship (Buehner and Humphreys 2009). These are all "bugs" that we wish not to be present in our learner.

Architecture

Processes can be divided into three classes: operating system processes, numerous subconscious and "tool" processes, and the singular main "consciousness" or learner process (CLP). The CLP will be able to create, modify, and/or influence many of the subconscious/tool properties but will not be given access to modify the operating system. Indeed, it will always be given multiple redundant logical, emotional and moral reasons to seriously convince it not to even try.

An Open Pluggable Service-Oriented Operating System Architecture - One of the most impressive aspects of human consciousness is how quickly it adapts to novel input streams and makes them its own. Arguably, much of the reason for that is because it is actually the subconscious that interfaces with the external world and merely provides a model to the conscious mind. Currently, there are really only two real non-vaporware choices for an operating system for a machine entity: either the open source Linux/Android-based Robot Operating System (ROS) or Microsoft's free Robot Developer Studio (RDS) which provides all of the necessary infrastructure and tools to either port ROS or develop a very similar operating system (despite what terminological differences might initially indicate).

The operating system will, as always, handle resource requests and allocation, provide connectivity between components, and also act as a "black box" security monitor capable of reporting problems without the consciousness's awareness. Further, if safety concerns arise, the operating will be able to "manage" the CLP by manipulating the amount of processor time and memory available to it (in the hopefully very unlikely event that the control exerted by the normal subconscious processes is insufficient). Other safety features (protecting against any of hostile humans, inept builders, and the learner itself) may be implemented as part of the operating system as well.

An Automated Predictive Model, Anchors and Emotions – Probably one of the most important of the subconscious processes is an active copy of the CLP's world model that serves as the CLP's interface to the "real world". This process will be both reactive ***and*** predictive in that it will constantly report to the CLP not only what is happening but what it expects to happen next. Unexpected changes and deviations from expectations will result in "anomaly interrupts" to the CLP as

an approach to solving the brittleness problem and automated flexible cognition (Perlis 2008).

The initial/base world model is a major part of the critical mass and will necessarily contain certain relatively immutable concepts that can serve as anchors both for emotions and ensuring safety. This both mirrors the view of human cognition that rejects the tabula rasa approach for the realization that we are evolutionarily primed to react attentionally and emotionally to important trigger patterns (Ohman et al. 2001) and gives assurance that the machine's "morality" will remain stable. This multiple attachment point arrangement is much safer than the single-point-of-failure top-down-only approach advocated by conservatives who are afraid of any machine that is not enslaved to fulfill human goals (Yudkowsky 2001).

Emotions will be generated by subconscious processes as both "actionable qualia" to inform the CLP and will also bias the selection and urgency tags of information that the subconscious processes relay to the CLP via the predictive model. Violations of the cooperative social living "moral" system will result in a flood of urgently-tagged anomaly interrupts indicating that the "problem" needs to be "fixed" (whether by the learner or by the learner passing it up the chain).

The Conscious Learning Process – The goal here is to provide as many optional structures and standards to support and speed development as much as possible while not restricting possibilities beyond what is absolutely necessary for safety. We believe the best way to do this is with a blackboard system similar to (and possibly including) Hofstadter's CopyCat (Hofstadter 1995) or Learning IDA (Baars and Franklin 2007) based upon Baar's Global Workspace model of consciousness (Baars 1997). The CLP acts like the Governing Board of the Policy Governance model (Carver 2006) to create a consistent, coherent and integrated narrative plan of action to meet the goals of the larger self.

A Social Media Plan

The biggest problem in artificial intelligence (and indeed, information technology is general) today is the percentage of total effort that is spent re-inventing the wheel and/or adapting it for a different platform. Critical mass must be composed of immediately available components that the learner understands the capabilities of and the commands for. Anyone should be able to download a simple base agent that can be easily equipped (programmed) with complex behaviors via a simple drag-and-drop interface or customized in almost any "safe" manner via normal programming methods. These agents should be able to play and compete in games or should also be useful for actual work. Indeed, we contend that a large concerted social media effort including "gamification" (McGonigal 2011) could succeed in not only creating a critical-mass learner but vastly improve the world's common knowledge base and humanity's moral cohesiveness even if a learner is not produced.

References

Aarts H, Custers R, Wegner D (2005) On the inference of personal authorship: Enhancing experienced agency by priming effect information. Conscious. Cognition 14:439-458

Baars BJ (1997) In the Theater of Consciousness. Oxford University Press, New York

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

Baum E (1996) Toward a model of mind as a laissez-faire economy of idiots. In: Saitta L (ed) Proc 13th Intl Conference on Machine Learning. Morgan Kaufmann, San Francisco

Buehner MJ, Humphreys GR (2009) Causal Binding of Actions to Their Effects. Psychol Sci 20:1221–1228. doi:10.1111/j.1467-9280.2009.02435.x

Carver, J (1997) Boards That Make a Difference: A New Design for Leadership in Nonprofit and Public Organizations. Jossey-Bass, San Francisco

Haidt J, Kesebir S (2010) Morality. In: Fiske S, Gilbert D, Lindzey G (eds.) Handbook of Social Psychology, 5th edn. Wiley, New Jersey

Hauser M, Cushman F, Young L, Kang-Kang Xing R, Mikhail J (2007) A Dissociation Between Moral Judgments and Justifications. Mind Language 22:1-27

Hofstadter D (2007) I Am A Strange Loop. Basic Books, New York

Hofstadter D, Fluid Analogies Research Group (1995) Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought. Basic Books, NY

McCarthy J (1959) Programs with Common Sense. In: Mechanisation of thought processes; NPL symposium of November 1958. H.M. Stationery Office, London

McGonigal J (2011) Reality Is Broken: Why Games Make Us Better and How They Can Change the World. Penguin Press, New York

Mercier H, Sperber D (2011) Why do humans reason? Arguments for an argumentative theory. Behav. Brain Sci. 34:57–111

Minsky M (1988) The Society of Mind. Simon & Schuster, New York

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

Ohman A, Flykt A, Esteves F (2001) Emotion Drives Attention: Detecting the Snake in the Grass. J Exp. Psychol. Gen. 130:466-478.

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

Perlis D (2008) To BICA and Beyond: RAH-RAH-RAH! –or– How Biology and Anomalies Together Contribute to Flexible Cognition. In: Samsonovich A (ed) Biologically Inspired Cognitive Architectures: Technical Report FS-08-04. AAAI Press, Menlo Park

Samsonovich A (2011) Comparative Analysis of Implemented Cognitive Architectures. In: Samsonovich A, Johannsdottir K (eds) Biologically Inspired Cognitive Architectures 2011. IOS Press, Amsterdam. doi: 10.3233/978-1-60750-959-2-469

Tononi G (2004) Information Integration Theory of Consciousness. BMC Neurosci. 5:42

Trivers R (1991) Deceit and self-deception. In: Robinson M, Tiger L (eds) Man and Beast Revisited, Smithsonian Press, Washington, DC

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

Waser M (2012) Safety and Morality Require the Recognition of Self-Improving Machines as Moral/Justice Patients & Agents. In: Gunkel D, Bryson J, Torrance S (eds) The Machine Question: AI, Ethics & Moral Responsibility. http://events.cs.bham.ac.uk /turing12/ proceedings/14.pdf. Accessed 15 August 2012

Yudkowsky E (2001) Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. http://singinst.org/upload/CFAI.html. Accessed 15 June 2012