What Is Artificial General Intelligence?
Clarifying The Goal For Engineering And Evaluation

Mark R. Waser

Books International
22883 Quicksilver Drive, Dulles, VA 20166
MWaser@BooksIntl.com

Abstract

Artificial general intelligence (AGI) has no consensus definition but everyone believes that they will recognize it when it appears. Unfortunately, in reality, there is great debate over specific examples that range the gamut from exact human brain simulations to infinitely capable systems. Indeed, it has even been argued whether specific instances of humanity are truly generally intelligent. Lack of a consensus definition seriously hampers effective discussion, design, development, and evaluation of generally intelligent systems. We will address this by proposing a goal for AGI, rigorously defining one specific class of general intelligence architecture that fulfills this goal that a number of the currently active AGI projects appear to be converging towards, and presenting a simplified view intended to promote new research in order to facilitate the creation of a safe artificial general intelligence.

Classifying Artificial Intelligence

Defining and redefining "Artificial Intelligence" (AI) has become a perennial academic exercise so it shouldn't be surprising that "Artificial General Intelligence" is now undergoing exactly the same fate. Pei Wang addressed this problem (Wang 2008) by dividing the definitions of AI into five broad classes based upon on how a given artificial intelligence would be similar to human intelligence: in structure, in behavior, in capability, in function, or in principle. Wang states that

> These working definitions of AI are all valid, in the sense that each of them corresponds to a description of the human intelligence at a certain level of abstraction, and sets a precise research goal, which is achievable to various extents. Each of them is also fruitful, in the sense that it has guided the research to produce results with intellectual and practical values. On the other hand, these working definitions are different, since they set different goals, require different methods, produce different results, and evaluate progress according to different criteria.

We contend that replacing the fourth level of abstraction (Functional-AI) with "similarity of architecture of mind (as opposed to brain)" and altering its boundary with the fifth would greatly improve the accuracy and usability this scheme for AGI. Since Stan Franklin proposed (Franklin 2007) that his LIDA architecture was "ideally suited to provide a working ontology that would allow for the discussion, design, and comparison of AGI systems" since it implemented and fleshed out a number of psychological and neuroscience theories of cognition and since the feasibility of this claim was quickly demonstrated when Franklin and the principals involved in NARS (Wang 2006), Novamente (Looks, Goertzel and Pennachin 2004), and Cognitive Constructor (Samsonovitch et. al. 2008) put together a comparative treatment of their four systems based upon that architecture (Franklin et al. 2007), we would place all of those systems in the new category.

Making these changes leaves three classes based upon different levels of architecture, with Structure-AI equating to brain architecture and Principle-AI equating to the architecture of problem-solving, and two classes based upon emergent properties, behavior and capability. However, it must be noted that both of Wang's examples of the behavioral category have moved to more of an architectural approach with Wang noting the migration of Soar (Lehman, Laird and Rosenbloom 2006; Laird 2008) and the recent combination of the symbolic system ACT-R (Anderson and Lebiere 1998, Anderson et al. 2004) with the connectionist [L]eabra (O'Reilly, and Munakata 2000), to produce SAL (Lebiere et al. 2008) as the [S]ynthesis of [A]CT-R and [L]ibra. Further, the capability category contains only examples of "Narrow AI" and Cyc (Lenat 1995) that arguably belongs to the Principle-AI category.

Viewing them this way, we must argue vehemently with Wang's contentions that "these five trails lead to different summits, rather than to the same one", or that "to mix them together in one project is not a good idea." To accept these arguments is analogous to resigning ourselves to being blind men who will attempt only to engineer an example of elephantness by focusing solely on a single view of elephantness, to the exclusion of all other views and to the extent of throwing out valuable information. While we certainly agree with the observations that "Many current AI projects have no clearly specified research goal, and

people working on them often swing between different definitions of intelligence" and that this "causes inconsistency in the criteria of design and evaluation", we believe that the solution is to maintain a single goal-oriented focus on one particular definition while drawing clues and inspiration from all of the others.

What Is The Goal of AGI?

Thus far, we have classified intelligence and thus the goals of AI by three different levels of abstraction of architecture (i.e. what it is), how it behaves, and what it can do. Amazingly enough, what we haven't chosen as a goal is what we want it to do. AGI researchers should be examining their own reasons for creating AGI both in terms of their own goals in creating AGI and the goals that they intend to pass on and have the AGI implement. Determining and codifying these goals would enable us to finally knowing the direction in which we are headed.

It has been our observation that, at the most abstract level, there are two primary views of the potential goals of an AGI, one positive and one negative. The positive view generally seems to regard intelligence as a universal problem-solver and expects an AGI to contribute to solving the problems of the world. The negative view sees the power of intelligence and fears that humanity will be one of the problems that is solved. More than anything else, we need an AGI that will not be inimical to human beings or our chosen way of life.

Eliezer Yudkowsky claims (Yudkowsky 2004) that the only way to sufficiently mitigate the risk to humanity is to ensure that machines always have an explicit and inalterable top-level goal to fulfill the "perfected" goals of humanity, his Coherent Extrapolated Volition or CEV. We believe, however, that humanity is so endlessly diverse that we will **never** find a coherent, non-conflicting set of ordered goals. On the other hand, the presence of functioning human society makes it clear that we should be able to find some common ground that we can all co-exist with.

We contend that it is the overly abstract Principle-AI view of intelligence as "just" a problem-solver that is the true source of risk and that re-introducing more similarity with humans can cleanly avoid it. For example, Frans de Waal, the noted primatologist, points out (de Waal 2006) that any zoologist would classify humans as ***obligatorily gregarious*** since we "come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy". If we, therefore, extended the definition of intelligence to "The ability **and desire** to live and work together in an inclusive community to solve problems and improve life for all" there would be no existential risk to humans or anyone else.

We have previous argued (Waser 2008) that acting ethically is an attractor in the state space of intelligent behavior for goal-driven systems and that humans are basically moral and that deviations from ethical behavior on the part of humans are merely the result of shortcomings in our own foresight and intelligence. As pointed out by James Q. Wilson (Wilson 1993), the real questions about human behaviors are not why we are so bad but "how and why most of us, most of the time, restrain our basic appetites for food, status, and sex within legal limits, and expect others to do the same."

Of course, extending the definition of intelligence in this way should also impact the view of our stated goal for AGI that we should promote. The goal of AGI cannot ethically be to produce slaves to solve the problems of the world but must be to create companions with differing capabilities and desires who will journey with us to create a better world.

Ethics, Language, and Mind

The first advantage of this new goal is that the study of human ethical motivations and ethical behavior rapidly leads us into very rich territory regarding the details in architecture of the mind required for such motivations and behaviors. As mentioned repeatedly by Noam Chomsky but first detailed in depth by John Rawls (Rawls 1971), the study of morality is highly analogous to the study of language since we have an innate moral faculty with operative principles that cannot be expressed in much the same way we have an innate language faculty with the same attributes. Chomsky transformed the study of language and mind by claiming (Chomsky 1986) that human beings are endowed with an innate program for language acquisition and developing a series of questions and fundamental distinctions. Chomsky and the community of linguists working within this framework have provided us with an exceptionally clear and compelling model of how such a cognitive faculty can be studied.

As pointed out by Marc Hauser (Hauser 2006; Hauser, Young and Cushman 2008), both language and morality are cognitive systems that can be characterized in terms of principles or rules that can construct or generate an unlimited number and variety of representations. Both can be viewed as being configurable by parameters that alter the behavior of the system without altering the system itself and a theory of moral cognition would greatly benefit from drawing on parts of the terminology and theoretical apparatus of Chomsky's Universal Grammar.

Particularly relevant for the development of AGI, is their view that it is entirely likely that language is a mind-internal computational system that evolved for internal thought and planning and only later was co-opted for communication. Steven Pinker argues (Pinker 2007) that studying cross-cultural constants in language can provide insight into both our internal representation system and when we switch from one model to another. Hauser's studies showing that language dramatically affects our moral perceptions argues that they both use the same underlying computational system and that studying cross-cultural moral constants could not only answer what is moral but how we think and possibly even why we talk.

Finally, the facts that both seem to be genetically endowed but socially conditioned and that we can watch the formation and growth of each mean that they can provide windows for observing autogeny in action.

Growing A Mind

One difference between most AGI researchers and many others working in the field of AI is the recognition that a full-blown intelligence is not going to be coded into existence. While AI researchers universally recognize the requirement of learning, there frequently isn't the recognition that the shortest path to AGI is to start with a certain minimal seed and to have the AGI grow itself from there. Indeed, many AGI research projects seem to have also lost this critical focus and be concentrating more on whether specific capabilities can be programmed in specific ways or on specific knowledge representations rather than focusing on the far more difficult subjects of what is required for effective growth from such a seed and how it might be implemented.

The interesting and important question, of course, is "What is the minimum critical mass for the seed AGI and what proportion of that mass is composed of hard-coded initial information as opposed to instructions for reasoning and growth?" Undoubtedly, there are many correct answers that will lead to a variety of different AGIs but we would prefer to pick one with a shorter path and time frame and a lesser amount of effort rather than a longer or more difficult path.

Daniel Oblinger (Oblinger 2008) has gone so far as to posit that it is possible that the majority of the work currently being done is unnecessary and can, and quite possibly will, be avoided by working instead on the bootstrapping process itself. It is his hope that a very minimal embedded system with the familiar AGI cognitive cycle (perceive/abstract/act or sense/cognize/act), the appropriate internal "emotional" drivers, and certain minimal social abilities will be able to use "embodiment scaffolding" and "social scaffolding" as a framework for growth that will permit the bootstrapping of strong performance from repeated iterations of weak learning. Both Marvin Minsky (Minsky 2006) and J. Storrs Hall (Hall 2007) give plausible models that we should be able to extend further.

On the other hand, longitudinal studies of twins raised apart (Bouchard 1990) show surprisingly high correlation levels in an incredible variety of choices, behaviors and outcomes. This, plus the examples of language and morality, suggests that much more of the details of intelligence are programmed in genetics than we might otherwise generally believe. It is our contention that studying the formation and growth of these examples will not only give us additional insight into the architecture of the human mind but is actually the quickest and most likely path to AGI by providing enough information to build the seed for a human-like architecture.

Architecture of Mind

Since we have previously noted that LIDA architecture implements and fleshes out a number of psychological and neuroscience theories of cognition and has already been deemed as an acceptable basis for comparison by the principals of a number of projects, we will consider it the consensus architecture of mind. The most salient features of LIDA's architecture are its cognitive cycle; the fact that it is very much an attentional architecture based upon Sloman's architecture for a human-like agent (Sloman 1999); and its use of Baar's global workspace theory of consciousness (Baars 1993, 1997, 2003; Newman, Baars and Cho 2003; Baars and Franklin 2007).

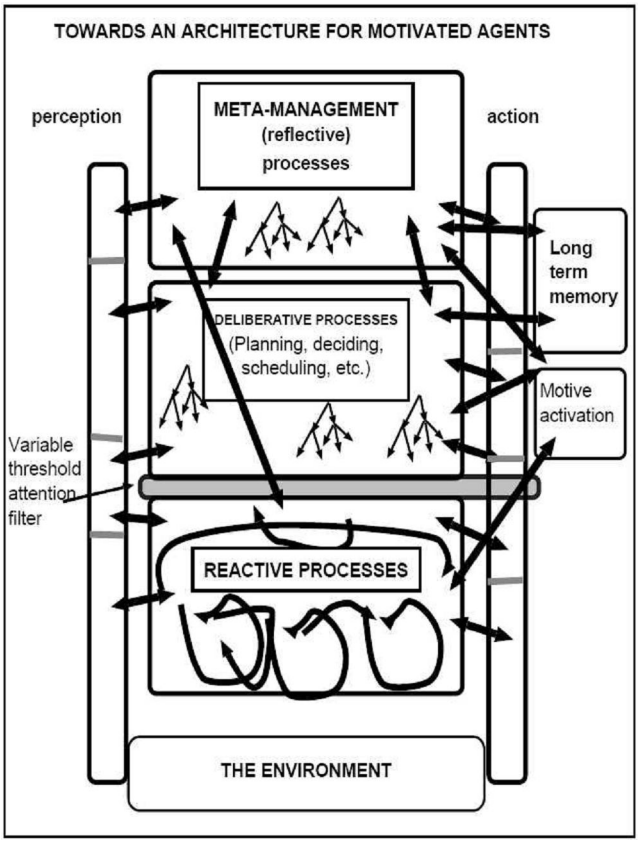

Figure 1. Sloman's human-like agent architecture

Franklin starts with an embodied autonomous agent that senses its environment and acts on it, over time, in pursuit of its own agenda. While it doesn't have the bootstrap view of what is the most minimal cycle that can build the simplest tool that can then be used as a building block to create the next tool, the LIDA model does include automization, the process of going from consciously learning something like driving to the effortless, frequently unconscious, automatic actions of an experienced driver. Since it is embodied and all cognitive symbols are ultimately grounded in perception, it is not subject to the symbol-grounding problem (Harnad 1990).

Franklin characterizes the simplicity of the initial agent by saying:

> It must have sensors with which to sense, it must have effectors with which to act, and it must have primitive motivators … [drives] … which motivate its actions. Without motivation, the agent wouldn't do anything. Sensors, effectors, and drives are primitives which must be built into, or evolved into, any agent.

Unfortunately, we would argue, for the purposes of both autogeny and morality, far too little attention has been paid to drives and their implementation.

Conscious Attention

In many ways, the most important feature of Sloman's architecture is the grey bar across the middle between conscious attentional processes and unconscious processes. Alfred North Whitehead claimed, "Civilization advances by extending the number of important operations which we can perform without thinking about them." We contend that the same is true of intelligence and would argue that there has been far less attention to the distinction between conscious and unconscious processing than we believe is warranted.

Experimental studies (Soon et. al. 2008) show that many decisions are made by the unconscious mind up to 10 seconds before the conscious mind is aware of it. Further, a study of the "deliberation-without-attention" effect (Dijksterhuis et al. 2006) shows clearly that engaging in a thorough conscious deliberation is only advantageous for simple choices while choices in complex matters should be left to unconscious thought. This effect is attributed to the fact that a person can pay conscious attention to only a limited amount of information at once, which can lead to a focus on just a few factors and the loss of the bigger picture. Logically, constraint satisfaction or optimization would seem to be an operation that would be best implemented on a parallel architecture (the unconscious) with a serial post-process (consciousness) for evaluating and implementing the result -- and another serial post-post-process for evaluating the results of the implementation and learning from them). Arguably, from the experiments presented above, it is entirely possible that the conscious mind merely "set up" the problem and then runs it on an unconscious tool.

Attention is also particularly important since it facilitates a second aspect of behavior control. As Minsky points out (Minsky 2006), most of our drives have both a sensory control and an attentional control. Sex not only feels good and but sexual thoughts tend to grab our attention and try to take over. Similarly, pain hurts and can distract us enough to prevent us from thinking of anything else.

Baars Global Workspace Theory postulates (Baars 1997) that most of cognition is implemented by a multitude of relatively small, local, special purpose processes, that are almost always unconscious. Coalitions of these processes compete for conscious attention (access to a limited capacity global workspace) that then serves as an integration point that allows us to deal with novel or challenging situations that cannot be dealt with efficiently, or at all, by local, routine unconscious processes. Indeed, Don Perlis argues (Perlis 2008) that Rational Anomaly Handling is "the missing link between all our fancy idiot-savant software and human-level performance."

A More Abstract View

An interesting abstraction of this architecture yields a simple view of intelligence, composed of just three parts, which is still complex enough to serve as a foundation to guide research into both the original evolution of the mind and also how individual human minds grow from infancy. The first part of the mind is the simple unconscious processes. Initially these must be hard-wired by genetics. The next part is a world model that has expectations of the world and recognizes anomalies. Desires are also a part of this world model. The third part is the integrative conscious processes that are not only invoked to handle anomalies but are also used to improve the world model and develop new unconscious processes.

This simple model captures many of the features of the human mind that many current models do not. Most important is the balance of the conscious processes being a slave to the desires and context of the world model formed initially and constantly revised by the subconscious yet being able to modify that model and create new subconscious processes. This is the dynamic of the seed that we contend is the quickest and safest path to AGI.

An important first question for ontogeny is where genetically "hard-coded" processes and model features stop and learned processes and features start. For example, evolution clearly has "primed" us with certain conceptual templates, particularly those of potential dangers like snakes and spiders (Ohman, Flykt and Esteves 2001). Equally interesting is the demonstration of the beginning of moral concepts like fairness in dogs (Range et al 2008) and monkeys (Brosnan and de Wall 2003).

What we believe to be most important, however, is further research into the development of a sense of self including its incredible plasticity in the world model and it's effects upon both the conscious and subconscious. Too many AGI researchers are simply waiting for a sense of self to emerge while the "Rubber Hand Illusion" (Botvinick and Cohen 1998) and the "Illusion of Body Swapping" (Petkova and Ehrsson 2008) give important clues as to how incredibly disparate subconscious processes will appear to the conscious mind merely as extensions to itself.

This is important point because it means that anything that can be plugged into the global workspace is immediately usable whether the conscious mind understands its internal operation or not. Of course, this immediately begs the question of exactly what the detailed "plug-and-play" interface specifications of the workspace architecture are – and this is where current systems all

differ. NARS uses Narsese, the fairly simple yet robust knowledge representation language of the system as an integration point. Novamente uses complex node-and-link hypergraphs. Polyscheme (Cassimatis 2005, 2006) uses numerous different representation schemes and attempts to implement the basic cognitive algorithms over them all.

More important than the knowledge representation scheme, we believe, however, is how the mechanism of attention is actually implemented. In LIDA, attention is the work of attention codelets that form coalitions to compete **in parallel** for access to the global workspace. Filtering occurs in multiple locations and is pretty much ubiquitous during cognition. Other systems merely label the various units of their representation schemes with interest values and priorities but there are tremendously variable degrees as to where attention falls on the spectrum of serial to parallel. It is our fear that the systems that do not dramatically limit the size of consciousness have deviated far enough from the model of human intelligence as to be in uncharted waters but only time will tell.

Conclusion

We have argued that creating an **Ethical Autogenous Attentional Artificial General Intelligence** (EA3GI) is likely to be the fastest and safest path to developing machine intelligence and that focusing on creating companions with differing capabilities and desires who will journey with us to create a better world instead of producing slaves to solve the problems of the world should be the consensus goal of AGI research.

References

Anderson, J.R.; Bothell, D.; Byrne, M.D.; Douglass, S.; Lebiere, C. and Qin, Y. 2004. An integrated theory of Mind. In *Psychological Review* 111:4.

Anderson, J.R. and Lebiere, C. 1998. *The atomic components of thought*. Mahwah, New Jersey: Erlbaum.

Baars, B.J. 1993. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B.J. 1997. *In The Theater of Consciousness: The Workspace of the Mind*. New York, New York: Oxford University Press.

Baars, B.J. 2003. How Does a Serial, Integrated, and Very Limited Stream of Consciousness Emerge from a Nervous System That Is Mostly Unconscious, Distributed, Parallel, and of Enormous Capacity? In Baars, B.J.; Banks, W.P.; and Newman, J.B. eds *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press.

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

Beck, J.; Ma, W.J.; Kiani, R; Hanks, T.; Churchland, A.K.; Roitman, J.; Shadlen, M.; Latham, P.E.; and Pouget, A. 2008. Probabilistic Population Codes for Bayesian Decision Making. *Neuron* 60(6): 1142 - 1152.

Botvinick, M. and Cohen, J. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391: 756–756.

Bouchard, T.J. Jr; Lykken. D.T.; McGue, M.; Segal, N.L.; and Tellegen, A. 1990. Sources of human psychological differences: the Minnesota Study of Twins Reared Apart. *Science* 250: 223–228.

Brosnan, S. and de Wall, F. 2003. Monkeys reject unequal pay. *Nature* 425: 297-299.

Cassimatis, N. 2005. Integrating Cognitive Models Based on Different Computational Methods. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. MahWah, New Jersey: Erlbaum.

Cassimatis, N. 2006. A Cognitive Substrate For Human-Level Intelligence. In Artificial Intelligence Magazine: 27. Menlo Park, CA: AAAI Press.

Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York, NY: Praeger Publishers.

Dijksterhuis, A.; Bos, M.; Nordgren, L.; and Baaren, R. van 2006 On Making the Right Choice: The Deliberation-Without-Attention Effect. *Science* 311: 1005 – 1007.

Franklin, S. 2007. A Foundational Architecture for Artificial General Intelligence. In Goertzel, B and Wang, P. eds. *Advances in Artificial General Intelligence*. Amsterdam, The Netherlands: IOS Press.

Franklin, S.; Goertzel, B.; Samsonovich, A. and Wang, P. 2007. Four Contemporary AGI Designs: A Comparative Treatment. In Goertzel, B and Wang, P. eds. *Advances in Artificial General Intelligence*. Amsterdam, The Netherlands: IOS Press.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42: 335-346.

Hall, J. 2007. Beyond AI: Creating the Conscience of the Machine. Amherst, NY: Prometheus Books.

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

Hauser, M. et al. 2007. A Dissociation Between Moral Judgments and Justifications. *Mind&Language* 22(1):1-27.

Hauser, M; Young, Y. and Cushman, F. 2008. Reviving Rawls' Linguistic Analogy: Operative principles and the causal structure of moral actions. In Sinnott-Armstrong ed. Moral Psychology and Biology. New York, NY: OUP.

Laird, J. 2008. Extending the Soar Cognitive Architecture. In *AGI 2008: Proceedings of the First AGI Conference*. Amsterdam, The Netherlands: IOS Press.

Lebiere, C.; O'Reilly, R.; Jilk, D.; Taatgen, N. and Anderson, J.R. 2008. The SAL Integrated Cognitive Architecture. In *AAAI Technical Report FS-08-04*. Menlo Park, CA: AAAI Press.

Lehman, J.; Laird, J. and Rosenbloom, P. 2006. *A Gentle Introduction To Soar, An Architecture For Human Cognition: 2006 Update*. Available at http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/GentleIntroduction-2006.pdf.

Lenat, D.B. 1995. Cyc: a large-scale investment in Knowledge Infrastructure. *Communications of the ACM* 38(11): 33-38.

Looks, M.; Goertzel, B. and Pennachin, C. 2004. Novamente: An Integrative Architecture for General Intelligence. In *AAAI Technical Report FS-04-01*. Menlo Park, CA: AAAI Press.

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

Newman, J.; Baars, B.J.; Cho, S.B. 2003. A Neural Global Workspace Model for Conscious Attention. In Baars, B.J.; Banks, W.P.; and Newman, J.B. eds *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press.

Oblinger, D. 2008. Towards an Adaptive Intelligent Agent. In *AAAI Technical Report FS-08-04*. Menlo Park, CA: AAAI Press.

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

O'Reilly, R.C. and Munakata, Y. 2000. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

Perlis, D. 2008. To BICA and Beyond: RAH-RAH-RAH! –or– How Biology and Anomalies Together Contribute to Flexible Cognition. In *AAAI Technical Report FS-08-04*. Menlo Park, CA: AAAI Press.

Petkova, V.I. and Ehrsson, H.H. 2008. If I Were You: Perceptual Illusion of Body Swapping. *PLoS ONE* 3(12): e3832.

Pinker, S. 2007. The Stuff of Thought: Language as a Window into Human Nature. New York, NY: Viking/Penguin Group.

Range, F.; Horn, L.; Viranyi, Z.; and Huber, L. 2008. The absence of reward induces inequity inversion in dogs. *Proceedings of the National Academy of Sciences USA* 2008 : 0810957105v1-pnas.0810957105.

Rawls, J. 1971. *A Theory of Justice*. Harvard Univ. Press.

Samsonovich, A.; De Jong, K.; Kitsantas, A.; Peters, E.; Dabbagh, N. and Kalbfleisch, M.L. 2008. Cognitive Constructor: An Intelligent Tutoring System Based on a Biologically Inspired Cognitive Architecture (BICA). In *AGI 2008: Proceedings of the First AGI Conference*. Amsterdam, The Netherlands: IOS Press.

Sloman, A. 1999. What Sort of Architecture is Required for a Human-like Agent? In Wooldridge, M. and Rao, A.S. eds *Foundations of Rational Agency*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Soon, C.S.; Brass, M.; Heinze, H-J; and Haynes, J-D. 2008. Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11: 543-545.

Tomasello, M. 2008. *Origins of Human Communication*. Cambridge, MA: MIT Press.

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

Waller, N.G; Kojetin, B.A.; Bouchard, T.J.; Lykken, D.T.; and Tellegen, A. 1990. Genetic and environmental influences on religious interests, attitudes, and values: a study of twins reared apart and together. *Psychological Science* 1(2): 138-142.

Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence*. Dordrecht, the Netherlands: Springer.

Wang, P. 2008. What Do You Mean By "AI"? In *AGI 2008: Proceedings of the First AGI Conference*. Amsterdam, The Netherlands: IOS Press.

Waser, M. 2008. Discovering The Foundations Of A Universal System Of Ethics As A Road To Safe Artificial Intelligence. In *AAAI Technical Report FS-08-04*. Menlo Park, CA: AAAI Press.

Yudkowsky, E. 2004. *Coherent Extrapolated Volition*. Available at http://www.singinst.org/upload/CEV.html.