The Evaluation of AGI Systems

Pei Wang

Temple University, Philadelphia, USA
http://www.cis.temple.edu/∼pwang/

Abstract

The paper surveys the evaluation approaches used in AGI research, and argues that the proper way of evaluation is to combine empirical comparison with human intelligence and theoretical analysis of the assumptions and implications of the AGI models.

Approaches of Evaluation

In recent years, the problem of evaluation is getting more and more attention in the field of Artificial General Intelligence, or AGI (GB09; LIIL09; LGW09; MAP$^+$09; Was09). Though the evaluation of research results is important in any field of scientific research, the problem has special difficulty in the current context of AGI, since the research activities belong to many different paradigms, and there seems to be no "neutral" way to compare them (Wan08).

In traditional AI research, since the projects are typically problem-oriented, it is natural to compare competing theories and techniques by the scope, correctness, and complexity of the algorithms involved, or by their performance on a set of benchmark instances of the problem. Obviously, this methodology is no longer applicable to AGI, given its stress of generality. Very commonly, one technique performs better on one problem than another technique, but not as well on a second problem. In this case, how can we decide which technique is "generally better"?

One solution proposed in mainstream AI is to select typical intellectual problems as "Grand Challenges" (Coh05; Bra06). Though such activities do stimulate interesting research, it still has the danger of leading the research to problem-specific solutions, no matter how carefully the problems are selected — after all, this was why problems like theorem proving and game playing were selected in the early days of AI, and the resulting techniques have not been generalized to other fields very well.

An alternative is to use multiple tasks (GB09) as a kind of "Cognitive Decathlon" (MJMH07). This approach clearly covers a larger field than a single challenge problem, but the selection of the components for the compound problem, as well as the way to calculate the "total score", may still look more or less arbitrary.

One way to be less problem-specific is to move away from testing problems, and to evaluate the systems according to certain properties (Min61; BBI$^+$98; LIIL09). While providing important insights, this approach also needs to justify its selection of desiderata and its method of overall evaluation, especially since the proposed property lists are all different from each other, and few technique has all the properties.

Due to the lack of a common evaluation methodology, in papers surveying AGI what we usually find are descriptions of the special properties (both desired and undesired) of each system, without an overall grading or ranking (PG07; DOP08). Though this kind of description is often fair and informative, it does not resolve the evaluation problem, but avoids it to a large extent. For the field of AGI as a whole, there is a need for clear, justified, and widely applicable ways to evaluate research results, rather than treating them as equally good (though different).

Obviously, an evaluation should start from the research goal — without a clearly specified destination, it is simply impossible to compare who is closer to it. Unfortunately, here the lacking of a common goal is exactly the reason for the lacking of a common evaluation methodology. By comparing two AAAI Presidential Addresses a decade apart (Dav98; Bra06), it seems clear that mainstream AI as a whole is not moving closer to a consensus on what its research goal really is. This diversity in objective inevitably causes the research efforts to move in different directions (Wan08), and to apply different evaluation criteria.

At the most general level, AI has been driven by the motivations of understanding human intelligence, and reproducing it in computers to solve practical problems. Therefore, there are two fundamental approaches for evaluation:

The empirical approach: To evaluate the intelligence of a computer system according to its similarity to human intelligence.

The theoretical approach: To evaluate the intelligence of a computer system according to its agreement with a theory of intelligence.

Roughly speaking, they correspond to the "think/act

like human" and "think/act rationally" in (RN02), respectively; in terms of the five types of "AI" defined in (Wan08), the "Principle-AI" is associated more closely with theoretical evaluation, while the other four with empirical evaluation, in different ways.

In the rest of the paper, the two approaches are discussed, compared, and related to each other.

Empirical Evaluation

Since the best known form of intelligence is human intelligence, it seems obvious that the ultimate criterion used in the evaluation of AI should be how close a system is to human intelligence. Concretely, for a given AI system, "empirical evaluation" means to test the system in various situations, and then its "level of intelligence", or some kind of "IQ", is measured by how close its input/output data is to human data in similar situations, just like how a theory in natural science is usually evaluated.

However, in the context of AGI, the problem is much more complicated. Even if we all agree that AI can be evaluated using "human data", there is still the question of *which* data should be used for this purpose. A brief survey of the related research shows that human intelligence has been studied at very different levels:

- There are researchers working on *brain* modeling, guided by the belief that the key of intelligence is hidden in the neural structure (dG07; HB04). However, even among them, there is still a huge difference on the type of data from neuroscience that is taken into consideration in the model.

- Starting from Turing (Tur50), there is the school that believes what really matters for intelligence are *behaviors*, and the intelligence of a computer system should be evaluated according to how the system acts like a human being. Similarly, in this school there are different opinions on what kind of behavior should be considered, and various variations of the Turing Test have been proposed (BBF01; Har00; MAP+09).

- There are psychologically motivated models of intelligence and cognition that focus on the *architecture* of the system, which is believed to be responsible for the production of the psychological data (AL98; Bac09; Fra07; New90). Among these architectures, again there are differences on which type of psychological data is considered.

- Many people judge the level of intelligence by the *problems* the system can solve, rather than by the details of the solving process. According to this kind of opinion, "general" (or "human-level") intelligence means the capability of solving a wide range of problems as human beings (McC07). Therefore, to evaluate the intelligence of an AI system, we can test it with examinations used in primary and secondary schools (GB09), or to see how many human jobs it can be employed in to replace a human (Nil05).

There are very different answers to the "which data?" question, because human intelligence is a complicated phenomenon that has been studied at many different levels of description, and with focus on different aspects. All of these research paradigms are valid and valuable, but they lead the research to different directions (Wan08), and inevitably require different evaluation criteria.

The above situation may remind us of the well-know story of the blind men and an elephant (Was09). Why not to take all types of human data into consideration altogether when evaluating a model?

One reason is that each description of human intelligence is carried out at a specific "level", using a vocabulary with specific scope and granularity. Consequently, it cannot be perfectly reduced or summarized into another level below or above. It is simply impossible to get a "complete description" of human intelligence that can satisfy all intellectual and practical purposes.

Even though in principle a computer can simulate any process in any accuracy, to duplicate human intelligence in this way is still an impossible practice, because to do that it is not enough to duplicate the neural electrical mechanism (since the other biological and chemical processes in the brain may be relevant), the complete brain (since the human body plays a significant role in cognition), or even a whole human being (since the influence of experience cannot be ignored in human behavior). Some people may argue that we can give the simulated human a simulated human experience in a simulated human world, but even in that scenario, the simulated world must be separated from our world, because we are not going to take a simulated human as a human, which means the simulated human cannot get our actual social experience.

Furthermore, even if such a complete simulation of human intelligence can be obtained, it is not what AI aims at. "Artificial Intelligence" is never an attempt to duplicate human intelligence *as it is*, in all aspects. Instead, it is the attempt to reproduce intelligence in computers, which are fundamentally different from the human brain at the hardware level. AI comes from the belief that "intelligence" is a phenomenon that may appear in various forms, and human intelligence is just one form of it. If computer intelligence can only be achieved via simulating human intelligence in all perceivable details, then the above belief will actually be falsified, rather than verified. Also, if "intelligence" merely means "human intelligence', no room will be left for possibilities like "animal intelligence", "group intelligence", or "extra-terrestrial intelligence". That would be a highly anthropocentric definition of intelligence. For the AI Dream to be fully realized, what to be created is a form of intelligence that is *similar* to human intelligence in certain essential aspects, but not in all aspects.

According to the above analysis, though it is possible, even necessary in certain sense, to evaluate AGI systems empirically by comparing them with human in-

telligence, the evaluation needs to be guided and justified by certain general principles, which decide the type of data to be considered in the evaluation. Though the human brain and mind is the major source of inspiration of AI research, it is not reasonable to judge the intelligence of an AGI system according to arbitrarily assembled human data.

Therefore, an empirical evaluation itself also needs justification and evaluation, and this "meta-evaluation" cannot be empirical, but must be theoretical. Before the behavior or performance of AGI systems are compared with certain human data, reasons must be given for why the type, scope, and granularity of the data are considered as directly related to aspects of human intelligence that can be meaningfully extended into machines, rather than as come from accidental factors that have little relevance to non-human intelligence. This is the case, because unlike models developed for pure biological, psychological, or evolutionary purposes, for AI/AGI models to be "just like human" may not be the ultimate aim.

Theoretical Evaluation

To many people, "intelligence" is the ability to find solutions that are correct and optimal in a certain sense; AI is the attempt to make computers do so, especially on problems that have not been solved in this sense; AGI means to achieve it in various domains by the same system.

With such an understanding of intelligence, the precondition for AI to be realized is a *theory* which specifies what is the correct or optimal solution to a problem, and how to find or build it. Such a theory must be general enough to be applied to various domains, and concrete enough to be implemented in a computer. Technically, it means the theory can be formalized. Clearly, it is a *normative theory* (like mathematics and computer science, which specify what should happen in a system to be created) of intelligence, rather than a *descriptive theory* (like biology and psychology, which specify what actually happens in a system to be understood).

The theoretical approach of evaluation means to compare the design and performance with the requirements and predictions of such a theory.

Though the above statement sounds reasonable, a question naturally follows: which theory? Currently there is no widely accepted "theory of intelligence". Instead, researchers have been building systems on different theoretical foundations:

- One of the most influential traditions in AI is based on *mathematical logic* (Hay77; McC88; Nil91). In the current AGI research, there are also many projects where the basic ideas are to extend classical logic in various directions (Bri08; GKSK09; Pol08).

- Given the intrinsically uncertain nature of many problems in AI, probability theory has been proposed as a proper theoretical foundation (Che85; Pea88), which has been accepted by more and more

people in recent years. It leads to the belief that an AGI system, or at least a large part of it, should be designed using probability theory and statistical methods (GIGH08; Mil08).

- Since AI systems are eventually implemented in computers, there is no surprise that many works are based on theory of computation, by analyzing problem-solving processes in terms of algorithm, computability, and computational complexity (Mar77; HF95). The same methodology is often used in the design and discussion of AGI systems (Bau04; Sch07).

- There have been various attempts to develop new normative theories specially for intelligence (Alb91; Hut05; Kug04; Rus97; Wan06). Each of the theories has some similarity with the traditional theories, but also introduces new postulates and assumptions, according to the beliefs of the researcher.

Since different theories usually give different evaluation results, once again we are facing a "meta-evaluation" problem: before AGI projects can be evaluated according to a theory, the theory itself needs to be analyzed and evaluated.

For our current purpose, a normative theory should be evaluated in two aspects: its intrinsic merit and its applicability to the AGI problem. Such a theory is based on a set of postulates and assumptions, from which the theorems and conclusions are derived by justified rules and procedures. Since the traditional theories (mathematical logic, probability theory, and theory of computation) have been thoroughly studied, the major problem about them is not in themselves, but in whether they can solve the problem of AGI.

Probably few people will insist that one of the traditional theories, in its canonical form, is sufficient for AGI. Instead, every "traditional-theory-based" AGI project has more or less extended and/or modified the theory, so as to make its assumptions and requirements satisfiable, and their conclusions competent to solve the problems in AGI. For instance, classical logic has been modified in various ways to deal with many types of uncertainty; probability theory is usually supplemented with assumptions on independence of the variables and the type of the distribution; theory of computation is often applied with simplifications and approximations here or there.

In this process, an important issue is the validness of the modification — to what extent a theory can be modified while remaining itself? Usually, new assumptions can be added, as far as they do not conflict with the existing ones. However, this is often not enough, and some postulates and axioms of the theory may be dropped or modified. One representative case is the reasoning model Probabilistic Logic Network, or PLN (GIGH08). When introducing it, the authors declare that "while probability theory is the foundation of PLN, not all aspects of PLN are based strictly on probability theory" (GIGH08, page 4). For instance, since "it is too much to expect any severely resource-constrained intel-

ligence to be fully self-consistent" (GIGH08, page 53), the consistency axiom of probability theory is dropped, and the system may assign different probability values to the same conclusion when following different reasoning paths. Though the reason is quite understandable, its effect needs clarification — should PLN be referred to as a revision of a probabilistic model of reasoning, or a model that is merely similar to a probabilistic model? When an axiom of probability theory has been violated, are the theorems of the theory (such as Bayes') still valid? Why? Answers to these questions can help us to decide whether PLN can be validated by its relation with probability theory.

The applicability problem is also a subtle one. Every normative theory is an idealization, and its applications into concrete domains are almost always rough, rather than accurate. If one assumption of the theory cannot be fully satisfied by the actual situation in a domain, will the theory become completely irrelevant, or remain an idealized case to be approximated? Such a situation can be found in (LH07), which defines the concept of "universal intelligence" as the ability to provide optimal solutions to problems, formalized in a theoretical model described in (Hut05). One notable assumption of the model is that it requires unlimited time-space resources, and the authors "consider the addition of resource limitations to the definition of intelligence to be either superfluous, or wrong" (LH07). Though it is well known that people in AI have very different understandings about "intelligence", it is obvious that "optimization with resource restriction" and "optimization without resource restriction" lead to very different models, though they are all "optimal", in certain (different) sense. If no concrete intelligent system can have unlimited resources, to what extent can these systems be properly evaluated according to a standard based on the opposite assumption?

To take the theoretical approach in evaluating AGI systems, we first need a well-established normative theory, with clearly stated assumptions, and conclusions implied by the assumptions. Furthermore, the assumptions should be satisfied by human intelligence, which is not only the best example of intelligence, but also widely believed to be selected by the evolution process, so is optimal in certain sense. For the same reason, the theory should be adequate in explaining various features of human intelligence that are desired to be reproduced in computers. Such a task should be carried out by comparing the theory with the reality of human intelligence. When approximation and simplification are needed, they should not completely change the nature of the problem. Otherwise, the theory cannot be convincingly used in the evaluation of systems — no matter how excellent the theory is on its own, it may not be applicable to the problem of AGI.

Therefore, a theoretical evaluation itself also needs justification and evaluation, and this "meta-evaluation" cannot be theoretical, but must be empirical. Before the design or performance of AGI systems are compared with what is required or predicted by a theory, reasons must be given to argue that the theory is at least satisfied by human intelligence. Otherwise the theory should not be designated as a "theory of intelligence".

Conclusion and Application

To establish widely applicable evaluation criteria is an important task for the field of AGI. It will not only enable systems and projects to be compared, but also guide the research in correct directions. Though in the near future the field will continue to host multiple research paradigms (Wan08), it is nevertheless necessary to avoid misleading evaluation approaches.

Given the nature of the AGI problem, there is no "natural" or "self-evident" way to evaluate the research. Any evaluation standard needs to be justified to be proper for the task. An evaluation proposal may be well-motivated, but still leads the research to a undesired direction by making an improper demand.

Compared to selected practical problems or functional features, it is more justifiable to evaluate an AGI system empirically according to human data, or theoretically according to an optimization model. In either case, the "meta-evaluation" is more general, reliable, consistent, and nonarbitrary.

The empirical approach of evaluation takes an AGI system as a descriptive model of human intelligence, made at an abstract level so that it becomes implementable in computers. For this approach, the meta-evaluation should take the form of a theoretical analysis, to argue that the selected data not only capture regularities in *human* intelligence, but also in other forms of intelligence. The major challenge in this approach is to separate the "intelligence-general" aspects of human intelligence from the "human-specific" aspects. The most likely mistake here is to propose a highly anthropocentric standard for AGI, which, even if possible to be achieved, will limit our imagination and innovation, and restrict the research in unnecessary ways resulting in "Artificial Human Intelligence", rather than "Artificial (General) Intelligence".

The theoretical approach of evaluation takes an AGI system as a normative model of intelligence that captures the essence of human intelligence at an abstract level. For this approach, the meta-evaluation should take the form of an empirical justification of the assumptions of the model (that is, they are indeed satisfied by human intelligence) and the model's explanatory power (that is, it is accountable for the cognitive functions observed in human intelligence). The major challenge in this approach is to identify the basics of human intelligence and to express them in a computer-implementable way. The most likely mistake here is to propose a highly biased standard for AGI, which, even if possible to be achieved, will lack key characteristics of intelligence as we commonly know, and to lead the research on deviant paths resulting in "Artificially Designated Intelligence", rather than "Naturally Designated Intelligence".

Now we see that the empirical approach and theoretical approach of evaluation actually depend on each other for the meta-evaluation. No matter which approach is selected, the other one will also be needed, though the two will serve different purposes in the whole evaluation process.

To make the above conclusion more concrete, let us briefly see how it is applied to the evaluation of the author's own research project, NARS (Wan06).

NARS is based on the belief that "intelligence" is the capability of *adaptation with insufficient knowledge and resources*. This belief itself is justified empirically — the human mind does have such capability (MR92).

The theory of intelligence built on this belief is a normative one, that is, it specifies how an intelligent system should work, not restricted by the biological, psychological, or evolutionary details of human intelligence. It is a theory of optimization, in the sense that if a system has to live and work in an environment, where the future cannot be accurately predicted, and the system's time-space resources are usually in short supply, then the theory provides the best design for the system. All the major design decisions of NARS are justified by theoretical analysis with respect to this objective, rather than by duplicating the human counterparts as faithfully as possible.

This theory is different from the traditional theories (classical logic, probability theory, theory of computation, etc.), mainly because of the above basic assumption. Since none of the traditional theories was developed for the problem of general intelligence, they do not assume the necessity of adaptation, nor the insufficiency of knowledge and resources in all aspects. Because this assumption plays a fundamental role in the theory, the issue cannot be resolved by minor extensions and modifications. Consequently, the NARS theory of intelligence is not based on any previous theory, though it surely inherits many ideas from them, and still uses them for subproblems here or there.

Even though NARS is primarily evaluated by theoretical analysis, to compare the performance and properties of the system with "human data" still makes sense. Since the human mind is the solution found by evolution for the same problem, it is not a coincidence that a truly intelligent AI system should share many similar properties with the human mind. For example, it should depend on certain learning mechanisms to deal with the uncertain future, while managing its own resources to achieve the best overall efficiency. However, due to its fundamental non-human hardware and experience, there is no reason to expect NARS to reproduce the human data exactly. "How the human mind does it" is a source of inspiration for the design decisions, but not their direct justification. For a new theory under development, empirical testing also reveals implicit and hidden implications of the assumptions — if NARS does something fundamentally different from human beings, then an explanation will be required, though it does not necessarily lead to a revision of the model.

In summary, to evaluate AGI systems, we need to properly combine the empirical approach and the theoretical approach, so as to find an identity for AGI that is neither too close to human intelligence (to become *Artificial Human Intelligence*), nor too far away from it (to become *Artificially Designated Intelligence*).

References

John R. Anderson and Christian Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

James S. Albus. Outline for a theory of intelligence. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):473–509, 1991.

Joscha Bach. *Principles of Synthetic Intelligence PSI: An Architecture of Motivated Cognition*. Oxford University Press, Oxford, 2009.

Eric B. Baum. *What is Thought?* MIT Press, Cambridge, Massachusetts, 2004.

Selmer Bringsjord, Paul Bello, and David Ferrucci. Creativity, the Turing Test, and the (better) Lovelace Test. *Minds and Machines*, 11(1):3–27, 2001.

Rodney A. Brooks, Cynthia Breazeal, Robert Irie, Charles C. Kemp, Matthew Marjanovic, Brian Scassellati, and Matthew M. Williamson. Alternative essences of intelligence. In *Proceedings of the Fifteenth AAAI/IAAI Conference*, pages 961–968, 1998.

Ronald J. Brachman. (AA)AI — more than the sum of its parts, 2005 AAAI Presidential Address. *AI Magazine*, 27(4):19–34, 2006.

Selmer Bringsjord. The logicist manifesto: At long last let logic-based artificial intelligence become a field unto itself. *Journal of Applied Logic*, 6(4):502–525, 2008.

Peter Cheeseman. In defense of probability. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 1002–1009, 1985.

Paul R. Cohen. If not Turings Test, then what? *AI Magazine*, 26:61–67, 2005.

Randall Davis. What are intelligence? and why? 1996 AAAI Presidential Address. *AI Magazine*, 19(1):91–111, 1998.

Hugo de Garis. Artificial brains. In Ben Goertzel and Cassio Pennachin, editors, *Artificial General Intelligence*, pages 159–174. Springer, Berlin, 2007.

Włodzisław Duch, Richard Oentaryo, and Michel Pasquier. Cognitive architectures: where do we go from here? In *Proceedings of the First Conference on Artificial General Intelligence*, pages 122–136, 2008.

Stan Franklin. A foundational architecture for artificial general intelligence. In Ben Goertzel and Pei

Wang, editors, *Advance of Artificial General Intelligence*, pages 36–54. IOS Press, Amsterdam, 2007.

Ben Goertzel and Stephan Vladimir Bugaj. AGI Preschool: a framework for evaluating early-stage human-like AGIs. In *Proceedings of the Second Conference on Artificial General Intelligence*, pages 31–36, 2009.

Ben Goertzel, Matthew Iklé, Izabela Freire Goertzel, and Ari Heljakka. *Probabilistic Logic Networks: A Comprehensive Framework for Uncertain Inference.* Springer, New York, 2008.

Helmar Gust, Ulf Krumnack, Angela Schwering, and Kai-Uwe Kühnberger. The role of logic in AGI systems: towards a lingua franca for general intelligence. In *Proceedings of the Second Conference on Artificial General Intelligence*, pages 43–48, 2009.

Stevan Harnad. Minds, machines and Turing: the indistinguishability of indistinguishables. *Journal of Logic, Language, and Information*, 9:425–445, 2000.

Patrick J. Hayes. In defense of logic. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 559–565, 1977.

Jeff Hawkins and Sandra Blakeslee. *On Intelligence.* Times Books, New York, 2004.

Patrick Hayes and Kenneth Ford. Turing Test considered harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 972–977, 1995.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2005.

Peter Kugel. Toward a theory of intelligence. *Theoretical Computer Science*, 317(1-3):13–30, 2004.

Christian Lebiere, Cleotilde Gonzalez, and Walter Warwick. A comparative approach to understanding general intelligence: predicting cognitive performance in an open-ended dynamic task. In *Proceedings of the Second Conference on Artificial General Intelligence*, pages 103–107, 2009.

Shane Legg and Marcus Hutter. Universal intelligence: a definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.

John E. Laird, Robert E. Wray III, Robert P. Marinier III, and Pat Langley. Claims and challenges in evaluating human-level intelligent systems. In *Proceedings of the Second Conference on Artificial General Intelligence*, pages 91–96, 2009.

W. Joseph MacInnes, Blair C. Armstrong, Dwayne Pare, George S. Cree, and Steve Joordens. Everyones a critic: memory models and uses for an artificial Turing judge. In *Proceedings of the Second Conference on Artificial General Intelligence*, pages 132–137, 2009.

David Marr. Artificial intelligence: a personal view. *Artificial Intelligence*, 9:37–48, 1977.

John McCarthy. Mathematical logic in artificial intelligence. *Dædalus*, 117(1):297–311, 1988.

John McCarthy. From here to human-level AI. *Artificial Intelligence*, 171:1174–1182, 2007.

Brian Milch. Artificial general intelligence through large-scale, multimodal Bayesian learning. In *Proceedings of the First Conference on Artificial General Intelligence*, pages 248–255, 2008.

Marvin Minsky. Steps towards artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49:8–30, 1961.

Shane T. Mueller, Matt Jones, Brandon S. Minnery, and Julia M.H. Hiland. The BICA Cognitive Decathlon: A test suite for biologically-inspired cognitive agents. In *Proceedings of the Behavior Representation in Modeling and Simulation Conference*, 2007.

Douglas L. Medin and Brian H. Ross. *Cognitive Psychology.* Harcourt Brace Jovanovich, Fort Worth, 1992.

Allen Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, Massachusetts, 1990.

Nils J. Nilsson. Logic and artificial intelligence. *Artificial Intelligence*, 47:31–56, 1991.

Nils J. Nilsson. Human-level artificial intelligence? Be serious! *AI Magazine*, 26(4):68–75, 2005.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, San Mateo, California, 1988.

Cassio Pennachin and Ben Goertzel. Contemporary approaches to artificial general intelligence. In Ben Goertzel and Cassio Pennachin, editors, *Artificial General Intelligence*, pages 1–30. Springer, New York, 2007.

John Pollock. OSCAR: an architecture for generally intelligent agents. In *Proceedings of the First Conference on Artificial General Intelligence*, pages 275–286, 2008.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, Upper Saddle River, New Jersey, 2nd edition, 2002.

Stuart Russell. Rationality and intelligence. *Artificial Intelligence*, 94:57–77, 1997.

Jürgen Schmidhuber. The new AI: General & sound & relevant for physics. In Ben Goertzel and Cassio Pennachin, editors, *Artificial General Intelligence*, pages 175–198. Springer, Berlin, 2007.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX:433–460, 1950.

Pei Wang. *Rigid Flexibility: The Logic of Intelligence.* Springer, Dordrecht, 2006.

Pei Wang. What do you mean by "AI"? In *Proceedings of the First Conference on Artificial General Intelligence*, pages 362–373, 2008.

Mark Waser. What is artificial general intelligence? Clarifying the goal for engineering and evaluation. In *Proceedings of the Second Conference on Artificial General Intelligence*, pages 186–191, 2009.