What we might look for in an AGI benchmark

Brandon Rohrer
Sandia National Laboratories
Albuquerque, NM, USA

Abstract

A benchmark in the field of Artificial General Intelligence (AGI) would allow evaluation and comparison of the many computational intelligence algorithms that have been developed. In this paper I propose that an ideal benchmark would possess seven key characteristics: fitness, breadth, specificity, low cost, simplicity, range, and task focus.

Introduction

As researchers in artificial general intelligence (AGI), we are sometimes asked, "What are you trying to do?" and "How will you know when you've done it?" And collectively we are forced to answer that we don't yet know. (Wan08) This is not for lack of ideas or effort. A reading of Goertzel and Pennachin's book surveying a broad swath of current AGI research makes it clear that many have thought deeply about the question, (GCP07) but the breadth of our backgrounds and our richness of diversity makes consensus challenging. There have been calls for a technical roadmap (LA09; GAS09) and concrete benchmarks (DOP08). This paper is intended as a contribution to the ongoing benchmark development effort.

Choosing a measurement device for AGI, a benchmark, is the key to answering questions about our aims. A benchmark implies a goal and implicitly contains a success criterion. Benchmarks can focus the efforts of a community; for all its limitations the Turing Test (Tur50) provided a fixed target for an entire subculture of artificial intelligence (AI) researchers, providing them with a common frame of reference and a shared language for efficient communication. An AGI benchmark would allow various approaches to be directly compared, promoting both cooperation and competition, as was seen most recently in large alliances and stiff competition in the race to win the Netflix Prize (Net09). Selecting an appropriate benchmark may greatly accelerate progress in AGI research.

Unfortunately, the selection of a good benchmark is difficult. A closely related problem is found in the assessment of human intelligence. The problem of measuring intelligence in humans is far from solved. While a number of formal measures exist, such as IQ tests, educational grade point averages, and standardized test scores, their merits are hotly contested. There is no consensus as to whether they are measuring "intelligence," or even a generally accepted definition of the word itself. There are also informal measures of intelligence, such as publication count or Erdös number in academic communities. It can also be argued that success in some critical endeavor reflects fitness and is an indirect indicator of intelligence. Depending upon one's peer group, success at a critical endeavor may be represented by one's salary, number of Twitter followers, or World of Warcraft level. From a biological standpoint, intelligence may be indirectly measured by one's reproductive fitness: the number of one's children or sexual partners. Despite (or perhaps due to) the large number of people that have devoted effort to defining a single useful measure of general human intelligence, no consensus has been reached. One complicating factor is that we have a conflict of interest; we may occasionally be guilty of advocating intelligence benchmarks at which we are likely to excel, rather than those which are likely to be the most useful.

Given the historical difficulty in choosing human general intelligence benchmarks, do we have a chance of choosing a non-human intelligence benchmark? We share many of the same challenges. We are no closer to a single definition of the term "intelligence." There is a profusion of potential measures. And we also may be tempted to advocate benchmarks at which our own systems are likely to excel. If there is one lesson we may learn from the history of human intelligence assessment it is that full consensus may be too ambitious. Our ultimate goals may be better served by choosing several benchmarks that are useful to many of us, rather than waiting until we find a single benchmark that is embraced by all.

This is not to say that any benchmark will do. It will require care not to choose a poor one. For example, performance on non-monotonic reasoning tasks has been proposed as a benchmark for artificial reasoning systems. However, closer examination revealed that human performance on the task was not well characterized, resulting in a machine intelligence benchmark that

was poorly aligned to human intelligence. (EP93) Illogic in human performance is not uncommon. Occasionally in the assessment of risk and reward, humans can be outperformed by rats. (Mlo08) This is not completely surprising. Deductive logic and the expectation maximization are tasks at which computers have outperformed humans for some time. But this example specifically highlights the pitfalls associated with benchmark selection. A benchmark based on reward maximization could result in a scale in which machines progress from human-level intelligence to the intelligence of a rodent.

There have been a number of benchmarks of machine performance that could be considered intelligence measures of a very narrow sort. These include classification datasets for supervised and unsupervised machine learning algorithms, (AN07) some of which contain images. (GHP07) There are also standard simulations on which reinforcement learning (RL) algorithms can compare their performance with each other, such as MountainCar (Moo90) and CartPole (GS93). There are a number of autonomous robotics competitions, which are benchmarks in the sense that they allow quantitative comparisons to be made between robotic systems. These include the robot soccer tournaments RoboCup (The09) and FIRA (FIR09), the autonomous submarine competition of the AUVSI (AUV09), AAAI robot contests, and perhaps best known, DARPA's driverless navigation Grand Challenges (DAR07). These events have demonstrated that a well-defined challenge can mobilize a large mount of effort and resources (which can be encouraged even further by the addition of several million dollars in prize money).

In the remainder of this paper I will enumerate the characteristics that, in my view, are desirable in an AGI benchmark, and propose a benchmark that meets those requirements. It is my hope that this proposal stimulates further discussion on the topic and contributes to the rapid selection of a provisional machine intelligence measure.

Benchmark criteria

Desirable attributes for an AGI benchmark are summarized in Table 1 and discussed below.

Table 1: Characteristics of a useful AGI benchmark

Fitness	Success on the benchmark solves the right problem.
Breath	Success on the benchmark requires breadth of problem solving ability
Specificity	The benchmark produces a quantitative evaluation.
Low Cost	The benchmark is inexpensive to evaluate.
Simplicity	The benchmark is straightforward to describe.
Range	The benchmark may be applied to both primitive and advanced systems.
Task Focus	The benchmark is based on the performance of a task.

Fitness

A benchmark implies a goal. While it may not always state a goal explicitly, it serves as an optimization criterion, which the research community uses to evaluate and direct its collective efforts. A useful benchmark will accurately reflect the goals of those subscribing to it. This may seem too obvious to merit attention, but it is surprisingly easy to pick a benchmark that does not fit this requirement. One purely hypothetical example of this might be found in a corporate environment where health and safety are high priorities. In order to reflect the importance placed on employee well-being, the number of reported injuries might be a reasonable choice of a performance benchmark. However, the simplest way to excel on this benchmark is for no employee to perform any work, thus avoiding the possibility of injury. This benchmark fails because it does not represent all the goals of the community, such as survival of the company and employee job satisfaction. However, this particular company is to be applauded for looking past the most common single corporate benchmark: stock price.

An AGI benchmark should reflect the goals of the AGI community. This will be challenging because those goals have not yet been agreed upon, leaving us without a clear target. However there have been a number of specific ideas proposed. (GAS09) The process of benchmark selection may accelerate and sharpen that discussion.

Another possible benefit of choosing a benchmark is that it may actually free us up from trying to extrapolate the results of our research out to a 10 or 50 year goal. We may be able to choose a benchmark that defines a research direction and let the end result be an emergent property of the researchers in our community each performing a local optimization: maximization against the benchmark. This approach may actually be more appropriate than defining a specific long-term goal at the outset. The research process is inherently uncertain and unpredictable. Having an emergent end goal would require a good deal of confidence in the benchmark, but would allow us to make progress toward a final goal that is currently beyond our capacity to visualize or articulate.

Breadth

Goertzel, Arel and Scheutz (GAS09) argued strongly for breadth (a very large task space) and accessibility (the attribute of requiring no previous task-specific knowledge) in an AGI benchmark. These two criteria capture a common sense among AGI researchers that a "general" intelligence can solve a more general class of problems than its forbears, and that it is, in a sense,

cheating for this to be done through extensive knowledge engineering or specialized heuristics. Weng introduced a related notion of task breadth that he termed muddiness. (Wen05) The ability to perform a broad set of tasks is a necessary characteristic of any system aspiring to human level intelligence.

The matching of human capability was the essence of the Turing Test and most AGI goal descriptions have been in a similar vein. In approaching such an ambitious problem it has been common practice in artificial intelligence research to reduce the breadth of the tasks while keeping the goal of human-level performance. There are strong temptations to reduce breadth: narrowing the task space and introducing task-specific system knowledge can produce far more eye-catching results and garner more attention, particularly from funding sources. However, our experience now shows that human-level performance in a narrow area, such as medical diagnoses or playing chess, does not necessarily generalize to a broader task set. Instead, it appears that maintaining breadth will ultimately be the more productive way to approach our long term goals. Keeping benchmarks broad while incrementally increasing performance expectations mimics the process followed by evolution during the development of animal intelligence. It is possible that following this course will automatically prioritize our efforts, focusing them on the most fundamental problems first.

Specificity

A useful benchmark will provide some quantitative measure of a system's value or performance. The best known benchmark from AI, the Turing Test, provides only a binary valuation, pass or fail. A number of similar tests have been proposed that may come closer to capturing the goals of AGI: the Telerobotic Turing Test (GAS09), the Personal Turing Test (CF05), and the Total Turing Test (Har91). Of course a binary benchmark is of limited use if we wish to evaluate systems that are not near the threshold of success. Turing-type tests could be made finer-grained by calibrating them against typical humans of varying ages, rather than setting a single threshold at the performance level of a typical adult. This notion of *cognitive age* (DOP08) could be further extended by calibrating performance against that of other species, resulting in a *cognitive equivalent organism*. A finer-grained measure, rather than a threshold, allows AGI candidates in various stages of development to be compared and progress to be charted over time. It also takes the pressure off researchers to define and come to consensus on a technological roadmap for developing AGI. (GAS09) Instead researchers can let the benchmark drive development priorities. In each particular approach, whatever aspect of technology would have the greatest impact on that system's benchmarked performance, that is where they can focus their efforts. The community would not need to spend time debating whether visual object recognition or non-monotonic logic needs to be addressed most urgently.

Even more useful would be a benchmark that mapped performance onto a scalar or vector of continuous or finely discretized values. With an appropriate mapping, common distance metrics such as the L^2 norm could be used to rank, order, and describe disparities between multiple AGI candidates. It would still be possible to set a Turing threshold, but a numerical benchmark result would allow evaluation of AGI efforts that fall short of human performance, as well as of those that exceed it.

Low Cost

An ideal benchmark will not require an inordinate amount of time, money, power, or any other scarce resource to evaluate. In order to be useful as a measurement device, it must be practical to apply. Even if it were excellent in all other respects, an incomputable benchmark would be of no practical value.

By taking advantage of economies of scale, competitions have proven to be an efficient way to evaluate a large number of systems in a single event. The overhead of administering the task, constructing the apparatus, and judging the results is shared among all the teams. A benchmark may also be able to use a competition format to reduce its cost in this way.

Simplicity

While not a requirement, it would be desirable for a benchmark to be simple in the sense that it could be accurately and concisely communicated to someone with only a high school (secondary school) diploma. Although the full motivation and justification for the benchmark may be much more complex, the ability to condense the success metric into a brief tagline can do a great deal to promote understanding in the wider scientific and non-scientific communities. This is particularly relevant to potential customers and funding sources. It is much easier to sell an idea if it can be clearly communicated. Simplicity will also promote accurate representation in popular media coverage of AGI. If we are able to provide brief summaries of our goals in the form of a soundbite, we can keep the stories more accurate. Otherwise we risk the distortion and misrepresentation that can inadvertently accompany technical reporting in the popular media.

Range

The best benchmark would be applicable to systems at all stages of sophistication. It would produce meaningful results for systems that are rudimentary as well as for systems that equal or exceed human performance. As was suggested earlier, a benchmark with a wide range of applicability would provide a full roadmap for development, giving direction both for immediate next steps and pointing toward long-range goals. This would have the added benefit of countering critics who might

claim that the goals of AGI are out of reach. A wide-range benchmark would imply near term, concrete goals by which we could measure and report our successes.

Task Focus

The four previous criteria (specificity, low cost, simplicity, and range) point toward a tool-agnostic task-focused benchmark. A performance measure of this type would not explicitly favor any particular approach (connectionist, symbolic, hybrid, or otherwise) but would reward each system purely on its demonstrated merits.

It is uncommon to have a scientific community united and defined by the problem it is trying to solve. It is much more common to have a community built around the use of a single computational, methodological, or modeling tool. This can be useful; it ensures that everyone understands everyone else's work. In a tool-centric community there is a common language and a shared set of assumptions that results in highly efficient communication. It is also easier to define who "belongs". Anyone whose work looks too unusual or unfamiliar is probably using a novel approach and is therefore an outsider.

Despite these benefits, tool-based definition is a luxury the field of AGI can't afford. The last several decades have demonstrated that focus on isolated toolsets is not necessarily the ideal approach to general AI. Any single tool may have hidden inductive biases that, if unacknowledged, can color the interpretation of its results. (THB07) There are now many significant efforts to combine multiple tools, specifically across connectionist-symbolic lines, one of the most notable of which is the DUAL architecture. (Kok94) Although it will require more effort in both explaining our work to each other and in grasping unfamiliar approaches, adopting a methodologically agnostic view greatly increases the size of the net we are casting for solutions. It is also an inoculation against intellectual inbreeding and unexamined assumptions, the primary symptoms of "looking where the light is."

One of the strongest arguments for a tool-centered approach to AGI is the biological plausibility of certain tools. However, this has proven to be a very elastic criterion. For example, artificial neural networks are based on approximate models of some neural circuits, yet some question the biological plausibility of their function. (AA09) Conversely, algorithms with no obvious biological implementation, such as the A* search, can mimic gross aspects of some human behaviors. Our neuroanatomic knowledge is too sparse at this point to conclusively specify or rule out algorithms underlying cognition. Most often the biological plausibility argument serves as a Rorschach test, helping us to expose our technical biases. And although there is some philosophical disagreement on this point among AGI developers, it could be argued that if a machine successfully achieves human-level performance on a broad intelligence metric, the biological plausibility of the approach is irrelevant.

Biological fidelity is itself an alternative to a task-based benchmark. This is the goal of model-based approaches to AGI. For now, the qualitative nature of biological fidelity makes it an unsatisfying benchmark candidate. Although serious efforts to quantify it are underway (LGW09), they are not yet mature. Interestingly, the proposed framework for establishing biological fidelity is also task-based, with the objective of matching human performance substituted for performance maximization. But until biological fidelity is concretely defined, establishing it more easily takes the form of a legal argument than a scientific one, with no conclusive way to resolve differences of opinion. However, seeking computational insights through biomimicry has been the genesis of many of our current computing tools and will undoubtedly serve as an ever-richer source of inspiration as our understanding of the brain matures.

A task-based benchmark has the additional benefit of keeping claims and counterclaims about competing approaches accurate. Without a mutually accepted basis for comparison, researchers are put in a difficult position when attempting to draw distinctions between their work and that of others. We are often reduced to speculating about the ultimate capabilities and limitations of both our own and others' approaches, a subjective and non-scientific endeavor that is frustrating and can spark animosity. This is an inherently problematic process, as we naturally underestimate those tools with which we are least familiar and overestimate those which we know best, particularly if we helped create them.

It may be reasonably argued that a benchmark with a strong task focus would provide limited support for the development of theory and mathematical analysis. But this is not necessarily the case. Theory and analysis have consistently provided insights that have enhanced performance. The adoption of a task-based benchmark would not make irrelevant rigorous mathematical work on AGI. It would only provide extra motivation to keep such theories grounded. These efforts make powerful mathematical statements about the potential capabilities of inductive problem solvers and thus are highly relevant to AGI. (Hut05; Sch04; Sch09) However, two conditions must be met for these efforts to directly contribute to improving performance on a task-based benchmark. 1) Every mathematical representation of the world makes modeling assumptions. These assumptions must not neglect or distort essential characteristics of the system being modeled. And 2) results must be reducible to practice. If a universal problem solver is mathematically defined, but could not be built with finite resources or run in finite time, it may be of limited value in pursuing a task-based benchmark. Reduction to practice is also a good method to verify that condition 1) was met.

A counter argument could be made that the development of intelligence should center exclusively on ana-

lytical and mathematical problems rather than physical or low-level tasks. The reasoning might be that higher level analytic and cognitive functions are uniquely human and should therefore be the sole focus of any effort to develop human level AI. But the fact remains that whatever cognitive abilities humans have acquired, they were preceded by the phylogenetically more basic abilities used by all mammals to find food, avoid threats, and reproduce. For this reason, more basic tasks of perception and physical interaction should not be neglected in favor of tasks that are more symbolic in nature.

Conclusion

A set of criteria for evaluating AGI benchmarks is proposed in Table 1. This is not intended to be a final answer to how to select a benchmark. Rather it is presented in the spirit of the "straw man," an imperfect incarnation that invites criticism, suggestions for improvement, and counterproposals. It is hoped that these criteria will promote discussion throughout the community, inspiring new and improved proposals for benchmarks which in turn will bring us closer to achieving our goals by clarifying them.

Acknowledgments

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

References

T. Achler and E. Amir. Neuroscience and AI share the same elegant mathematical trap. In *Proc 2009 Conf on Artificial General Intelligence*, 2009.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

AUVSI. Auvsi unmanned systems online, 2009. http://www.auvsi.org/competitions/water.cfm, Accessed September 22, 2009.

R. Carpenter and J. Freeman. Computing machinery and the individual: The Personal Turing Test. Technical report, Jabberwacky, 2005. http://www.jabberwacky.com/personaltt, Accessed September 22, 2009.

DARPA. Darpa urban challenge, 2007. http://www.darpa.mil/grandchallenge/index.asp, Accessed September 22, 2009.

W. Duch, R. J. Oentaryo, and M. Pasquier. *Frontiers in Artificial Intelligence Applications*, volume 171, chapter Cognitive architectures: Where do we go from here?, pages 122–136. IOS Press, 2008.

R. Elio and F. J. Pelletier. Human benchmarks on AI's benchmark problems. In *Proc 15th Congress of the Cognitive Science Society*, pages 406–411, Boulder, CO, 1993.

FIRA. Federation of International Robosoccer Association Homepage, 2009. http://www.fira.net/, Accessed September 22, 2009.

B. Goertzel, I. Arel, and M. Scheutz. Toward a roadmap for human-level artificial general intelligence: Embedding HLAI systems in broad, approachable, physical or virtual contexts. Technical report, Artificial General Intelligence Roadmap Initiative, 2009. http://www.agi-roadmap.org/images/HLAIR.pdf. Accessed September 21, 2009.

B. Goertzel and Eds. C. Pennachin. *Artificial General Intelligence*. Springer, 2007.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. http://authors.library.caltech.edu/7694.

S. Geva and J. Sitte. A cart-pole experiment for trainable controllers. *IEEE Control Systems Magazine*, 13:40–51, 1993.

S. Harnad. Other bodies, other minds: A machine incarnation of an old philisophical problem. *Minds and Machines*, 1:43–54, 1991.

M. Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer-Verlag, Berlin Heidelberg, 2005.

B. N. Kokinov. The DUAL cognitive architechture: A hybrid multi-agent approach. In *Proceedings of the Eleventh European Conference on Artificial Intelligence*. John Wiley and Sons, 1994.

S. Livingston and I. Arel. AGI roadmap, 2009. http://agi-roadmap.org/, Accessed September 22, 2009.

C. Lebiere, C. Gonzales, and W. 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*. Atlantis Press, 2009.

L. Mlodinow. *The Drunkard's Walk: How Randomness Rules Our Lives, 8th Printing Edition*. Pantheon, 2008. See Chapter 1.

A. Moore. *Efficient Memory-Based Learning for Robot Control*. PhD thesis, University of Cambridge, 1990.

Netflix. Netflix prize homepage, 2009. http://www.netflixprize.com/, Accessed September 23, 2009.

J. Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54:211–254, 2004.

J. Schmidhuber. Ultimate cognition *à la* Gödel. *Cognitive Computing*, 1:177–193, 2009.

P. Tino, B. Hammer, and M. Bodén. *Perspectives of Neural-Symbolic Integration*, volume 77, chapter 5. Markovian bias of neural-based architectures with

feedback connections, pages 95–133. Springer-Verlag, Heidelberg, Germany, 2007.

The RoboCup Federation. RoboCup Homepage. http://www.robocup.org/, 2009. Accessed September 22, 2009.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

P. Wang. *Frontiers in Artificial Intelligence Applications*, volume 171, chapter What do you mean by AI?, pages 362–373. IOS Press, 2008.

J. Weng. Muddy tasks and the necessity of autonomous mental development. In *Proc. 2005 AAAI Spring Symposium Series, Developmental Robotics Symposium*, Stanford University, Mar 21-23 2005.