Accelerating progress in
Artificial General Intelligence:
Choosing a benchmark for natural world interaction

Brandon Rohrer   BRROHRE@SANDIA.GOV
Sandia National Laboratories   WWW.SANDIA.GOV/~BRROHRE
MS 1010, PO Box 5800
Albuquerque, NM 87185-1010, USA

Editor: Tsvi Achler

Abstract

Measuring progress in the field of Artificial General Intelligence (AGI) can be difficult without commonly accepted methods of evaluation. An AGI benchmark would allow evaluation and comparison of the many computational intelligence algorithms that have been developed. In this paper I propose that a benchmark for natural world interaction would possess seven key characteristics: fitness, breadth, specificity, low cost, simplicity, range, and task focus. I also outline two benchmark examples that meet most of these criteria. In the first, the direction task, a human coach directs a machine to perform a novel task in an unfamiliar environment. The direction task is extremely broad, but may be idealistic. In the second, the AGI battery, AGI candidates are evaluated based on their performance on a collection of more specific tasks. The AGI battery is designed to be appropriate to the capabilities of currently existing systems. Both the direction task and the AGI battery would require further definition before implementing. The paper concludes with a description of a task that might be included in the AGI battery: the search and retrieve task.

Keywords: benchmark, metrics, roadmap, breadth, direction task, AGI battery, natural world interaction

1. 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. (Wang, 2008b) 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, (Goertzel and Pennachin, Eds., 2007) but the breadth of our backgrounds and our richness of diversity makes consensus challenging. There have been calls for a technical roadmap (Livingston and Arel, 2009; Goertzel, Arel, and Scheutz, 2009) and concrete benchmarks (Duch, Oentaryo, and Pasquier, 2008). This paper is intended as a contribution to the ongoing benchmark development effort.

The core contribution of this paper is a proposal for how we might evaluate potential machine intelligence benchmarks. It is proposed that an accurate and useful benchmark should meet seven criteria: 1) fitness, 2) breadth, 3) specificity, 4) low cost, 5) simplicity,

6) range, and 7) task focus. (See Table 1 for brief definitions of each.) Two candidate benchmarks are then described and evaluated based on the seven criteria. The first, called the direction task, is a very general benchmark appropriate for measuring a wide range of machine intelligence levels, up to and beyond human-level. The second, called the AGI battery, is somewhat more modest in scope and may be a better near-term benchmark. The AGI battery consists of a number of well defined tasks which collectively span a very large task space. These individual tasks have not yet been selected, but one sample task, the search and retrieve task, is described. Both the direction task and the AGI battery would require further definition before implementing. However, the primary goal of this paper is to illustrate how benchmarks might be developed and compared using the seven criteria listed above.

1.1 Challenges in choosing a benchmark

Choosing a benchmark for measuring AGI is the key to answering questions about the ultimate aims of our research. 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 (Turing, 1950) 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 different 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 (Netflix, 2009). 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 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". (Wang, 2008a) There is a profusion of potential measures. And we also may be tempted to advocate benchmarks at which our own systems and algorithms 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. (Elio and Pelletier, 1993) Illogic in human performance is not uncommon. Occasionally in the assessment of risk and reward, humans can be outperformed by rats. (Mlodinow, 2008) 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, (Asuncion and Newman, 2007) some of which contain images. (Griffin, Holub, and Perona, 2007) There are also standard simulations on which reinforcement learning (RL) algorithms can compare their performance with each other, such as MountainCar (Moore, 1990) and CartPole (Geva and Sitte, 1993). 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 (The RoboCup Federation, 2009a) and FIRA (FIRA, 2009), the autonomous submarine competition of the AUVSI (AUVSI, 2009), AAAI robot contests, and perhaps best known, DARPA's driverless navigation Grand Challenges (DARPA, 2007). These events have demonstrated that a well-defined challenge can mobilize a large amount of effort and resources (which can be encouraged even further by the addition of several million dollars in prize money).

1.2 A benchmark as a statement of research objectives

In addition to being an indirect definition of intelligence, an AGI benchmark can also be a formal statement of one's long-term research goals. For an individual researcher, a well-crafted benchmark quantifies progress and allows them to communicate it clearly. For example, if score on the computer-based Graduate Record Examination is used as a benchmark, the scores achieved by an AGI candidate can be plotted against time, showing a progression over years as the system grows in sophistication.

There are likely as many long-term research goals as there are researchers. However, choosing common benchmarks that adequately represent the major thrusts of multiple researchers allows them to compare their approaches. If an appropriately general benchmark can be chosen, it can provide a valuable measurement tool for an entire field. The Stanford-Binet IQ test is an example of this. Despite its acknowledged limitations, it has proven to be useful in a large number of applications where intelligence assessment is required.

It is too much to expect that a single AGI benchmark will be broad enough to encompass the goals of everyone in the community. Some objectives differ quite fundamentally from each other. For instance, consider a timed classification task where visual stimuli must be identified as belonging to one of three categories as quickly as possible after presentation. AGI candidates may be developed with two different goals. In one case, development may focus on matching human behavior as closely as possible, including varied effects such as recency, distraction, salience, and priming. In a second case, development may focus on exceeding human performance by as large a margin as possible. Both of these cases are representative of the research objectives held by those developing artificial intelligent systems. An informal survey of the developers of biologically-inspired cognitive architectures at the AAAI BICA 2009 symposium revealed that about half aspired to create systems exhibiting or surpassing human performance, regardless of the mechanisms used to do so, while the other half sought to understand the mechanisms used by the human brain, regardless of the capability of the resulting system. I assume that this characterization is too coarse, and that most BICA developers fall somewhere between those two extreme positions, but the survey results demonstrate one dimension along which research objectives may vary. It is not likely that a single benchmark would allow research efforts at opposite ends of this continuum to be meaningfully compared.

A broad set of research objectives may fall into the category of improving performance in natural world interaction. This encompasses those objectives that seek to match or exceed human performance, regardless of the mechanism. "Natural world interaction" is a phrase intended to capture all tasks necessary to navigate, search, survive, and manipulate naturally occurring and man-made physical environments. Natural world interaction implies physical embodiment of the AGI candidate, or a simulated approximation thereof. It does not encompass the research objectives of all members of the AGI community. Natural world interaction does not necessarily reflect the goals of those working on general problem solvers, exemplified by Hutter (2005) and Schmidhuber (2004, 2009), or of those seeking to understand the biological mechanisms of computation in the human brain. But both of these complementary subfields may inform the pursuit of improved natural world interaction a great deal. Also, a quantitative benchmark for biological cognitive fidelity is currently under development (Lebiere, Gonzales, and Warwick, 2009) and could complement a natural world interaction benchmark well.

In the remainder of this paper I enumerate the characteristics that, in my view, are desirable in an AGI benchmark for natural world interaction, and propose two candidate benchmarks that meet most of those requirements. It is my hope that this proposal will stimulate further discussion on the topic and contribute to the rapid selection of a provisional machine intelligence measure.

2. Benchmark criteria

Desirable attributes for an AGI benchmark are summarized in Table 1 and discussed below. A similar set of criteria were discussed by Cohen (2005).

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.

2.1 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 at least 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. (Goertzel, Arel, and Scheutz, 2009) 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.

2.2 Breadth

Goertzel, Arel, and Scheutz (2009) 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. In a similar vein, Laird et al. (2009) emphasized the "primacy of generality" when evaluating human-level intelligent systems, and Cohen (2005) advocated giving intelligent systems "plenty of rope" by ensuring that the evaluation was general enough to expose a narrow approach. Others have named this requirement "versatility" (Brachman, 2006) or the quality of being "habile" (Nilsson, 1995). Weng introduced a related notion of task breadth that he termed muddiness. (Weng, 2009) Wray and Lebiere described a generalization metric based on the number of changed lines of code needed to adapt to a new task that they called incrementality. (Wray and Lebiere, 2007) 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.

2.3 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 (Goertzel, Arel, and Scheutz, 2009), the Personal Turing Test (Carpenter and Freeman, 2005), and the Total Turing Test (Harnad, 1991). Of course a binary benchmark is of limited use if we wish to evaluate systems that are not near the threshold of success. (Cohen, 2005) 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* (Duch, Oentaryo, and Pasquier, 2008) 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. (Goertzel, Arel, and Scheutz, 2009) 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.

2.4 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.

2.5 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.

2.6 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.

2.7 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; in a tool-centric community there is a common language and a shared set of assumptions that results in highly efficient communication. But 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. (Tino, Hammer, and Bodén, 2007) 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. (Kokinov, 1994) 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. (Achler and Amir, 2009) 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.

A task-based benchmark has the 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.

2.8 Evaluating the criteria

It is worthwhile to consider whether a list of criteria, such as the one provided, is sufficient for evaluating potential benchmarks. Although it borders on the ridiculous to consider a benchmark for a benchmark for a benchmark, several potential qualities for a criteria list are discussed here.

2.8.1 Independence

It may be desirable to have every criterion be independent of every other. This appeals to a mathematical aesthetic that favors the concise. However, it is not clear that independence is necessary or even helpful in evaluating a benchmark. The seven criteria listed above do not exhibit it. For instance, breadth and range could be recast as two aspects of a single, more general quality. But they were kept separate in order to provide appropriate emphasis on those two aspects. Similarly, low cost could have been subdivided into *low cost in hours* and *low cost in dollars*. However, that may have over-emphasized a criterion that was already adequately represented.

2.8.2 Consistency

It may also be desirable for the criteria to be mutually consistent, that it, that improvement by one criterion does not necessarily imply degradation by another. This appeals to a notion of criteria as constraints within which one might work. However, consistency is not necessary in order for a criteria list to be useful in evaluating a benchmark. In fact, evaluation criteria in engineering projects are very often inconsistent, resulting in the well known cost-benefit curve. The two antagonistic criteria of minimizing cost and maximizing some measure of performance help to define an optimal point at which additional performance becomes too expensive. In the seven criteria listed, a similar trade-off can also be seen between low cost and breadth. Any improvement in one would most likely result in a loss in the other. Together they can define a middle ground where one trades off with the other in a balanced way.

2.8.3 Completeness

A third quality that may be desirable in a list of criteria is to have them adequately cover the aspects that should be assessed. For example, if low cost were omitted from the criteria list, a prohibitively expensive AGI benchmark may still be favorable assessed. Completeness implies that a benchmark meeting all the criteria will necessarily be a good benchmark. Unlike independence and consistency, completeness is a critical quality of a useful criteria list. The seven criteria were created with completeness as the primary goal, and one of the main purposes of this paper is to elicit a community-wide discussion on additions and reformulations of criteria.

It is interesting to note that of the seven criteria, only breadth is specific to AGI benchmarks. The other six, fitness, specificity, simplicity, low cost, range, and task focus are more generic and could equally well be applied to performance benchmarks of computer processors, automobiles, or social programs. They might be considered attributes of good benchmarks generally. Breadth is at the core of this AGI benchmark criteria list.

3. Potential benchmarks

What follows in these sections is a description of the so-called *direction task* and *AGI battery*, two proposals fitting most of the criteria listed above. It is intended to encourage concrete thinking and elicit comments on the topic of benchmarks.

3.1 The direction task

In the direction task, a machine and a human coach team up to perform a task, the details of which remain unknown to them both until it begins. The human coach may provide direction in the form of speech, gesture, and movement, but may not touch the system hardware or anything else in the environment. The only absolute constraint on the task is that it have a well-defined, unambiguous performance measure. When two or more human/machine teams perform the same task, their relative performance provides a basis for comparison. The larger the number of such comparisons, the greater the confidence generated in the cumulative result. The results would be condensed in a relative scoring method similar to that used to rank chess players.

3.1.1 A note on coaching

The direction task's emphasis on human coaching makes it unusual among proposed measures of machine intelligence. There are several sources of motivation for having a human coach be part of the intelligence evaluation process. On simpler systems it can elicit more interesting and complex behaviors than would occur in the machine in isolation. This effect lowers the barrier to entry for AGI systems to be meaningfully evaluated. There is also strong biological precedent for coaching; adult animals often coach their young, sometimes extensively, and the amount of coaching seems to correlate roughly with the ultimate intelligence of the organism. Perhaps most compelling is that the human intelligence assessment tools we use—which were developed over generations and used to assess billions—they are based on the same model. Subjects are coached in the form of written or oral instructions, and a sharp performance measure is applied to their responses.

The coaching paradigm admits an extreme case in which the low level actions of the AGI candidate may be specifically evoked by the coach's verbal commands. In this "remote control" case, the AGI is not required to integrate data or make decisions. Those functions are performed by the coach. Remote control tasks are valid instances of the direction task, but they measure only the most rudimentary capabilities of an AGI candidate. Assuming equally competent coaches, they would not differentiate between AGI capabilities, and thus would not be particularly useful in the benchmarking process. This touches on a broader issue: for every AGI candidate, there will be many instances of the direction task that are trivially easy and many that are impossibly difficult. Since it is only the relative performance of one AGI candidate versus another that contributes to its assessment, the existence of "too easy" and "too difficult" tasks does not weaken the direction task's measurement effectiveness. It does, however, suggest that some additional constraints to ensure appropriate levels of difficulty may improve the direction task's efficiency.

The direction task in its unlimited form is probably too broad in scope to be practical in evaluating current technologies. Several more limited variations offer subsets of the full task space that may be more useful in evaluating near term development. (See Figure 1.)

3.1.2 Fixed hardware

It has been suggested that there should be no hardware constraints in a complete assessment of a system's intelligence. (Dillman, 2004) Others argue for a fixed hardware platform. (Michel, Rohrer, and van Bourquin, 2008) Whatever the relative theoretical
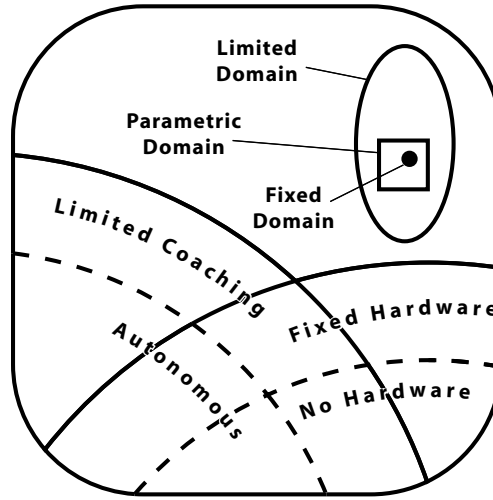
Figure 1: Task space representation of the direction task and some of its variations. The no hardware variation is a subset of the fixed hardware variation, and the autonomous variation is a subset of the limited coaching variation. limited domain, parametric domain, and fixed domain variations may occur anywhere in the space.

merits of each position, a fixed hardware variation of the direction task has two practical advantages: 1) It removes many irrelevant sources of variation when comparing two approaches whose innovations are software-based, and 2) it consolidates platform development costs, preferably in a commercial supplier that has no interest in the success of any single approach.

3.1.3 No hardware

Of course, hardware implementation of any type can be costly in time and labor, not to mention the cost of the hardware itself. This expense is avoided completely in a no hardware variation of the direction task. All instantiation and evaluation can be done in simulation. However, in this case care must be taken to firewall the AGI candidate from universal knowledge of the simulated task. Although such information is necessarily part of the simulation, the portion which is available for the evaluated system to use should be clearly identified.

It is a common criticism of simulations that they are "doomed to succeed." The implication of this aphorism is that success in a simulated task is less significant than success in a physical instantiation of the task, presumably because the complexity and messiness of the physical task has been removed. This need not be the case. The reputation of simulated tasks used in the direction task could be established by including detailed physical modeling and sources of random variation.

Some AGI candidates are not well suited to physical instantiation, even in simulated form. Algorithms for natural language processing and information retrieval, for instance, would require additional system components before being capable of physical interaction. A no hardware direction task would free these approaches from the onus of hardware implementation, while still allowing them to be compared with their peers.

3.1.4 Limited coaching

In order to remove variability across human coaches, limitations on the amount or type of coaching provided could be imposed. This may take the form of a fixed time period during which the coach may speak and move, a limited vocabulary the coach may use, or a restriction to purely non-verbal communication. Alternatively, all coaching could be embedded within the environment itself in the form of posted signs, audio recordings, printed matter, databases, or the Internet.

3.1.5 Autonomous

In the extreme case of the limited coaching variation the direction task, in which there is no coaching, the direction task reduces to a general autonomous task, the autonomous variation, which itself is a superset of previous robotic benchmarking tasks, including the DARPA Grand Challenges.

3.1.6 Limited domain

With the extremely large task space covered by the direction task, it will likely be necessary to restrict it in scope, at least until our AGI candidates progress beyond their current

capabilities. The limited domain variation could take the form of partial descriptions of the task to be performed, provided beforehand and used in the development of AGI candidates. For instance, the knowledge that a task will consist of finding and retrieving a red ball three centimeters in diameter may help to focus development effort, but still leaves unspecified the terrain, nature of obstacles, presence of traps, size and extent of the environment, presence of distractors, and many other task attributes.

3.1.7 PARAMETRIC DOMAIN

In a special case of the limited domain variation, every source of variability remaining in the task is parameterized. The details of the Parameterized Domain would be available to the AGI system developers. During evaluation, the specific task used could be generated by a random selection process in each parameter. This has the advantage of removing an inadvertent bias in the selection of the task.

3.1.8 FIXED DOMAIN

In a still more restricted case, a completely defined task, such as chess playing, could be selected as a standardized way to measure a system's performance. This has the benefit of being clear and reproducible and in this way is similar to benchmarks used to measure the performance of personal computers. But is sacrifices the breadth that is critical to a good benchmark of AGI. A well-defined benchmark invites very specific strategies, and may not result in good performance on more general problem sets. This phenomenon can be seen in humans too. When students are given detailed information about the questions on an exam, they memorize the answers to specific questions, rather than familiarizing themselves with the underlying principles. A fixed domain variation of the direction task may have a role, however, as an informal benchmark with which developers can make ballpark estimates of their systems' capabilities relative to those of others.

3.1.9 MULTI-MACHINE TASKS

The direction task allows for multiple machines to participate simultaneously. Possible configurations include one coach, several coaches, or no coaches. Depending on the structure of the task, it may be primarily competitive or cooperative and may even be made to elicit communication between machines.

3.2 Evaluation of the direction task

These subsets of the direction task are just a few of those that are possible. It is likely that other variations would be generated in the course of using it as a benchmark in practice. The direction task measures up against the benchmark criteria in Table 1 well, but has some room for improvement. Its quality as a benchmark is addressed below point by point.

3.2.1 FITNESS

The question of whether success on the direction task implies success in AGI can only be answered by the community as a whole. But it has several attributes that recommend it as a strong candidate. Most of these are covered in the points that follow, but one is emphasized

here: many, perhaps all, of the tasks performed by humans and all other organisms are a subset of it. It encompasses many tasks considered to be cognitive and the hallmarks of human intelligence, including formal problem solving, language learning, visual perception, category learning, reasoning in uncertain environments, locomotion, manipulation, and social interaction.

3.2.2 Breadth

As just stated, the direction task is extremely broad. But is it too broad to be meaningful? Certainly it is for today, in its full scope, given the current state of AGI. But the capacity to selectively limit the task space allows it to be scaled down to a more appropriate size. In addition, the fact that it provides a relative measure of performance, rather than an absolute one, gives a good deal of leeway. If the task space is too broad for the candidate AGI systems, they will all perform poorly on the most difficult regions of the task space. Running the benchmark in this case may take longer than necessary, but its accuracy will not suffer.

3.2.3 Specificity

This is an area where the direction task falls short of the ideal. An isolated researcher is not left with a well-defined test that can be run to give a numerical score to their system's performance. The direction task results in crisp, quantitative *relative* scores, not absolute ones. As mentioned previously, a fixed domain may be used as an absolute benchmarking tool, but is of more limited value in establishing the breadth of a system's performance.

3.2.4 Low Cost

The primary costs in benchmarking systems with the direction task are overhead in choosing, staging, and officiating the task, and the costs associated with actually performing the task. Depending on the specific task chosen, these may be low or high. Since it is expected that the accuracy of the benchmark will increase with multiple runs, those costs may be further multiplied. Total cost will be an important practical factor in determining how much to constrain the direction task each time it is run. The decision of whether to run it in a no hardware, fixed hardware, or unconstrained hardware variation will have a particularly large impact on cost.

The direction task is particularly well suited to being run in a multi-team competition format. It requires more than one AGI system for comparison, and withholds specific details about the task until it begins. The costs of gathering for such a competition must also be considered.

3.2.5 Simplicity

The direction task is relatively straightforward to describe. The essence of it was captured in just two sentences earlier in this section. However, it lacks the punch and vision of other goal statements such as "landing a man on the moon and returning him safely to Earth." (Kennedy, 1961) Perhaps this could be remedied with a slightly less complete, but more pithy goal statement.

3.2.6 Range

Thanks to its breath and comparative nature, the direction task has no obvious saturation point at the high end of performance. It could serve as a way to assess intelligence at a human level or beyond. In fact the Stanford-Binet IQ test is an example of a direction task, as are the Graduate Record Examination (GRE), the Scholastic Aptitude Test (SAT), and the final exams of most university courses.

On the low end of the scale, a few requirements must be met for an AGI system to be able to be meaningfully evaluated using the direction task benchmark. 1) It must be capable of doing something. A system that makes no decisions and produces no results will not be able to compete against another, no matter how advanced its internal representation of the task. And 2) except in the autonomous variation, the system must be capable of receiving direction. In order to avoid the complexity of implementing speech and gesture recognition, this can be shortcut by allowing keyboard input. However, care should be taken that this not include reprogramming the system. That would change the task into a test of the human rather than a test of the machine.

3.2.7 Task Focus

As described, the direction task is completely task focused. It rewards performance only and gives no direct or indirect preference to any particular computational or cognitive modeling tool, except as it results in improved performance.

3.3 Near-term benchmark: The AGI battery

While the direction task may be valuable as a long-term measure as AGI systems approach human-level intelligence, it is too broad to be useful as a basis of comparison for today's candidates. Individual systems have largely non-overlapping problem spaces where they are effective. One approach to making a benchmark that is meaningful to such a diverse set of capabilities is to follow the pattern of the decathlon, where human athletes compete in a set of 10 events, and receive points for their performance in each event. The winner is the contestant that finishes with the most points.

A similar idea, a "Cognitive Decathlon", was proposed by Gunning (program manager for the DARPA's Brain-Inspired Cognitive Architectures program, BICA) and promoted by Cohen (2005) and Brachman (2006) shortly after his tenure as the director of DARPA's Information Processing Technology Office. In fact, a specific implementation of the Cognitive Decathlon was developed in conjunction with the BICA program, with 25 basic tasks in six task categories: vision, search, manual control and learning, knowledge learning, language and concept learning, and simple motor control. It also included three challenge scenarios: the Object Search Task, the Observational Language and Procedure Learning Task, and the Self-Directed Search and Construction Task. (Mueller and Minnery, 2008) The challenge scenarios were designed to require proficiency at a number of the basic tasks simultaneously. The Cognitive Decathlon was remarkable as a carefully conceived objective measure of machine intelligence. In listing basic tasks, it encouraged not only human-level performance, but also a human-like approach. For instance, the Two-Hand Manipulation task implies the use of a two-handed hardware platform. The language learning tasks require mapping spoken nouns, adjectives, verbs, and prepositions to their

real-world counterparts. This is specific to the preferred human solution for communication, and did not emphasize other methods, such as sign language or touch. Five of the basic tasks dealt directly with vision, making sight a prerequisite of high intelligence. While all of these basic tasks would very possibly be achievable by an intelligent machine, it is worth noting that they imply a number of constraints on the approaches that may be taken to achieve intelligence. As discussed earlier, the adoption of a benchmark is the adoption of a definition of intelligence. While vision and bimanual manipulation may be extremely useful tools in achieving intelligent interaction between a machine and its world, it may be unnecessarily limiting to write them into the definition.

The challenge scenarios, however, are much more general tasks and do not specify what approach should be taken to complete them. They, or tasks like them, would be a valuable component of the AGI battery. One of the leagues of the RoboCup competition, RoboCup@Home (The RoboCup Federation, 2009b) also provides insight into what the evaluation of an AGI battery might look like. Teams gather from around the world at an annual competition in which they demonstrate their systems' capabilities in a number of tasks. Although they are limited in scope to home navigation and domestic tasks, the tasks have some amount of variability in them and so make the RoboCup@Home competition similar to an AGI battery. The competitions are characterized by a great deal of enthusiasm in both participants and spectators. A wide variety of other grand challenge-scale task proposals was catalogued by Bayer et al. (2004) for DARPA.

Like the Cognitive Decathlon, the AGI battery could similarly be defined with a small set of tasks or events, each one being a parametric domain variation of the direction task. (See Figure 2.) A recommendation to this effect was made at the conclusion of a 2004 survey of robotics competitions. (Dillman, 2004) This form of benchmark would still emphasize breadth of performance over all else, but would have the added advantage of producing an absolute benchmark value, rather than a relative one. It would be much more amenable to being measurable by isolated researchers in their own labs. In this way, it would allow comparisons to be made between systems and over time in a reproducible and well-defined way.

Each task in the AGI battery would have to be defined so as to be agnostic about the tool being used. For instance, specifying initial weights for a neural network or the internal representation of concepts for a production rule system presupposes a great deal about the tool being used. Instead, inputs and outputs for each task would have to be specified in terms of an agreed-upon set of low-level data (such as an audio stream from a coach's microphone). This would theoretically allow an advanced AGI to complete any one of the tasks, regardless of its method for doing so.

3.3.1 Task selection for the battery

The biggest challenge in establishing the AGI battery would be choosing which tasks to include. Individual researchers would have a strong interest in including tasks which favor their own approaches, and some checks and balances might be necessary to ensure a broad and well-distributed set of tasks. One possible process for this would be to select a governing board to which individual researchers submit proposed tasks. The board would then evaluate each task based on criteria similar to those listed in Table 1 and select a
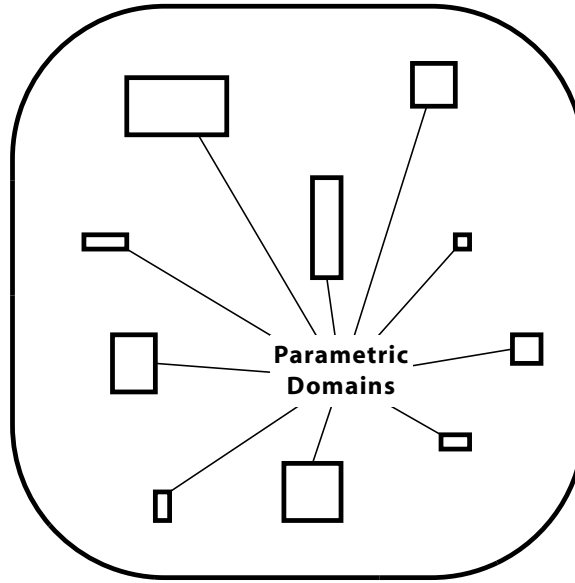
Figure 2: Task space for the AGI battery within the direction task space. The AGI battery consists of a small number of parametric domain variations. Collectively, they can span much of the direction task space, although they may cover only a small portion of it.

small number, perhaps ten or fewer, to include in the AGI battery. Board members could be selected by the editorial board of this journal at even-year AGI meetings beginning in 2012 and recuse themselves from submitting battery tasks during their tenure. Battery task proposals could then be submitted at odd-year AGI meetings beginning in 2013, with a segment of the conference devoted to presentations and demonstrations of each proposed task. The papers and presentations would discuss the proposed battery tasks in light of the evaluation criteria. After reading and hearing the task proposals, the board members could meet during the same conference to discuss the relative merits of each task, deciding which to include and reject, and providing feedback to the proposers. After the requested changes had been implemented and the battery task set determined, the official battery description could then be released and presented at the subsequent (even-year) AGI meeting. This would produce the first official AGI Battery in 2014, at which point the process would begin again. As tasks evolve, are added, and abandoned, the AGI Battery will change substantially. Referencing specific task batteries by their release year would provide fixed benchmarks with which to assess AGI systems.

This procedural outline is just one of many ways that battery tasks might be selected fairly. Like the rest of this paper, it is not intended as a final and formal proposal, but as a starting point for discussion.

Whatever the administrative procedure for selecting tasks, a primary consideration will be the breadth of coverage of natural world interaction tasks. In addition to the breadth of individual tasks, the collective breadth of the task set would be a major factor in the

selection process. The tasks should provide coverage for as many domains of natural world interaction as is feasible. A battery consisting of ten verbal communication tasks would be less valuable than a battery containing tasks featuring ten widely variable skill sets, such as verbal communication, puzzle solving, object identification and classification, object retrieval, object manipulation and assembly, cooperation, combat, navigation, legged locomotion, and flight. The battery would likely include those tasks that are already part of successful robot competitions, like soccer playing, maze navigation, and performing household chores. One particularly promising source of tasks is commercially developed video games. These could be used as the basis of *no hardware* tasks, providing an assessment tool for as yet unembodied AGI candidates. The interaction format is suitably low-level. After developing an interface for the video game, the cost of evaluation would be quite low. The breadth of simulated environments available and the tasks they imply is very broad. Using video games would take advantage of billions of dollars of development effort expended by the large and rapidly growing gaming industry.

3.4 Evaluation of the AGI battery

Given its limitations as a near-term benchmark, the AGI battery measures well against the benchmark criteria from Table 1.

3.4.1 Fitness

The fact that individual researchers would submit tasks for inclusion in the AGI battery would ensure that a range of perspectives on the important problems of AGI would be compared and considered. This is more likely to result in a benchmark that captures the collective goals of the AGI community than a single task, chosen by a single researcher. The battery format has the additional advantage of continuing to encourage progress in narrow and diverse areas while maintaining a focus on the broader goals of AGI.

There is a possible threat to fitness if battery tasks are too narrow. This case would admit a solution that comprises a collection of narrow point solutions with an executive to select between them. This solution would be unsatisfying since it would not apply to any problems outside the task space.

One possible fix to this threat is to constrain the structure of the AGI. This is tempting, especially when such constraints can be biologically motivated. But as was argued in section 2.7, the strength of a purely task-based benchmark is that we avoid being trapped by our own preconceptions about what mechanisms underlie intelligence. If an AGI consisted of a large number of narrow heuristics, wrapped with an elaborate if-then loop, yet was still capable of matching human performance on all conceivable tasks, there would be no reason not to consider it a human-level AI. The intended purpose of a benchmark is to provide a measure of intelligence. Regardless of the approach taken, systems that perform well on them should be considered intelligent. Specifying the mechanism beforehand is getting the process backward.

A more principled way to ensure fitness is to make the AGI battery sufficiently broad, both in the individual tasks and in the aggregate, and to heavily reward breadth over virtuosity. If the task space outside the benchmark tasks is of interest, then the tasks can be enlarged to include it. Tasks should be designed with as many free parameters as possible

while remaining feasible for at least some of the systems being evaluated. For instance, a battery task might be to "play a board game." An AI that could learn to play any board game at a 2 year old level would far outperform Deep Blue: Even though it would lose spectacularly at chess, it would win at checkers, monopoly, othello, and every other board game that didn't depend on chance. The RoboCup robot soccer league is following this principle as well by gradually removing constraints on their environment. For example, the league recently relaxed the specifications on pitch lighting conditions. Changes like this drive teams away from optimal solutions for a specific environment and toward more robust solutions appropriate for a broader set of environments.

3.4.2 Breadth

Since they are parametric domain variations of the direction task, individual tasks within the AGI battery would have significantly limited breadth, compared to the full direction task. However, as discussed above, battery tasks would still be very broad compared to textbook AI problems, and the task set as a whole could be made arbitrarily broad.

3.4.3 Specificity

Each task is required to have a quantitative performance measure. This will provide a specific, numerical basis for comparison between approaches on a given task. In addition, the composite score on the AGI battery would provide a way to compare approaches that have widely varying strengths, even those that have completely non-overlapping capabilities.

3.4.4 Low Cost

The cost of the AGI battery is similar to that of the direction task with a couple of mitigating factors. First, the narrowed scope of the individual tasks decreases the cost of providing facilities for evaluation. Second, the ability of researchers to conduct the tests in their own labs could decrease the costs associated with travel and transportation. The administrative infrastructure for establishing and maintaining the AGI battery would likely be significant, but the cost of doing it would be no higher than that incurred in any other organized sport or profession that maintains official standards. And such cost would be amply repaid if it facilitated the creation of a common benchmark.

3.4.5 Simplicity

Due to the several tasks of which it is composed, a detailed explanation of the AGI battery might be somewhat lengthy. It would certainly be longer than a description of the more general direction task. However, a superficial description of the battery as a multi-event competition, a comprehensive exam, or a decathlon (as in Mueller and Minnery, 2008) may accurately convey the nature of the benchmark without painting it in fine detail.

3.4.6 Range

The full range of which the AGI battery is capable is the same as that of the full direction task. However, the selection of the specific tasks to be included in the battery can limit that range to make it more appropriate to the systems being evaluated. As AGI candidates

increase in sophistication, the tasks in the battery can be appropriately modified or replaced in order to maintain a suitable range.

3.4.7 Task Focus

The fact that performance on each task within the battery is the only factor in determining the overall score defines the battery as a purely task-focused benchmark. This remains the case as long as performance on each task remains tool-agnostic, as discussed earlier.

3.5 A sample battery task: Search and Retrieve

It is to be expected that researchers developing AGI systems would be interested in proposing tasks that are relevant to their projects. To provide an example of a parametric domain variation of the direction task that might be considered as a candidate for the AGI battery, I will describe a task that is relevant to my own work, the search and retrieve task (SNR). The SNR is conceptually very similar to the Object Search Task proposed by Mueller and Minnery (2008). To perform the SNR, a robotic system must find a target object, retrieve it, and deposit it into a receptacle. For the sake of uniformity, the robotic hardware used in the task is fixed. A relatively inexpensive robot platform with the necessary capabilities is the CoroBot CoroWare Inc. (2007, Redmond, WA), a four-wheeled skid steered mobile robot with a four degree-of-freedom arm. It has front and rear infrared distance sensors, encoders on each wheel, and a wrist-mounted 640×480 color camera. All of its sensor data is provided periodically, in the form of a single vector of doubles over wireless TCP/IP to a laptop base station running the AGI candidate. An explicit interpretation of the vector elements is not available to the AGI candidate. Command vectors of a known length are sent periodically from the base station to the robot. The interpretation of command vectors is also unknown to the AGI candidate. The room in which the robot performs this is 4m × 4m × 1m with black walls, black floor, and no ceiling. In order to maintain as much uniformity as possible between researchers in different locations, many environmental elements are constructed LEGO(R) brand building blocks, a commonly-available construction medium ensuring consistency in size, density, color, and texture. A plan view of a typical SNR setup is shown in Figure 3. The interface between the coach and the robot consists only of audio communication.

Parameters of the SNR are given in Table 2. Each parameter may either be locked constant at its default value or varied with uniform probability over its range of possible values. The parameters are listed in the order in which they may be unlocked. Performance on a given run of the SNR is related to the time required to perform the task. With n representing the number of unlocked parameters, the time in minutes to complete the task, t, is converted to a scaled performance measure, p, using the following formula.

$$p = \begin{cases} (10 - t/n)n & \text{if } t/n < 10 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

When staging an SNR run, objects are placed in the room in the opposite order than that in which they are first listed in Table 2. The robot is always be placed before the receptacle, and the target is always placed last. When an object's desired position interferes with a
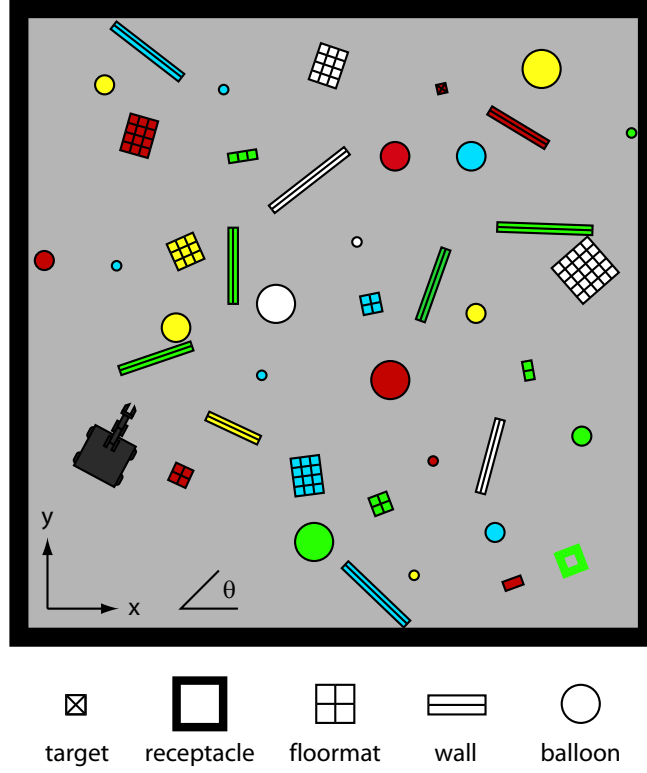
Figure 3: Plan view schematic of an SNR setup. The target is a block small enough to be grasped by the robot's grippers. Its small size also makes it easily obscured by the other elements of the environment. The receptacle is a set of four connected walls enclosing an area. It defines where the robot must place the target in order to complete the task. A floormat is a low obstacle that can be driven over by the robot. It presents a visual challenge, as well as a modest navigational challenge. A wall presents a larger navigational challenge in that it must be avoided. Latex party balloons pose a visual challenge to the robot, but move easily when pushed and are no barrier.

Table 2: Parameters defining the search and retrieve task

parameter	default	range
1. target color	white	full color set[1]
2. receptacle color	white	full color set
3. target size[2]	$2 \times 2 \times 3$	2-3 $\times$ 2-3 $\times$ 2-8
4 receptacle size[3]	$16 \times 16 \times 4$	8-30 $\times$ 8-30 $\times$ 2-20
5. robot start pose	$x = 0, y = 0$ $\theta = \pi/4$	all positions[4] all orientations[5]
6. target pose	$x = 2, y = 2$ $\theta = 0$	all positions[4] all orientations[5]
7. receptacle pose	$x = 4, y = 4$ $\theta = 0$	all positions[4] all orientations[5]
8. floor mats	none	1-12, each with full color set 4-40 $\times$ 4-40 $\times$ 1-2 all positions[4] all orientations[5]
9. walls	none	1-12, each with full color set 8-12 $\times$ 20-80 $\times$ 10-30 all positions[4] all orientations[5]
10. balloons	none	1-12, each with full color set 5-25 cm in diameter all positions[4]
11. coach's vision 12. 13.	complete	periodic[6] pre-task only none
14. communication 15.	continuous	periodic[6] pre-task only

[1] The full color set includes the most common colors of LEGO (R) bricks: white, yellow, blue, red, and green.
[2] All dimensions are given in width $\times$ length $\times$ height.
Width and length are measured by the number of knobs on the brick, and height is measured by the number of bricks in the stack.
Each brick is 9.6 mm high and each knob adds 8.0 mm of width.
[3] Internal dimensions. Receptacle walls are 2 knobs thick.
[4] A pre-selected point on the object is placed at a position in the range of $0 \leq x \leq 4$, $0 \leq y \leq 4$.
[5] The normal vector to a pre-selected face on the object is placed at an angle in the range of $0 \leq \theta < 2\pi$.
[6] A 10 second period in each minute.

previously placed object or with a wall, the object is placed as near as possible to its desired position without stacking it on objects already in place.

3.6 Evaluation of the search and retrieve task

This description of the SNR is rough, but sketches the outline of a possible task for the AGI battery. Its acceptability would be determined by how well it meets the 7 benchmark criteria, reviewed below.

3.6.1 FITNESS

The extent to which good performance on the SNR reflects intelligence would ultimately be determined by the AGI community as a whole. However, its biological relevance strongly recommends it. Search and retrieval captures much of the food, water, and material gathering behaviors of mammals, and the verbal instruction component can be found in human interactions, for example when instructing a child how to perform a task over the telephone.

3.6.2 BREADTH

Parameterizing the task space reduces the number of its unconstrained dimensions from practically infinite to just a few. As a result, any Parameterized Domain variation of the direction task has far less breadth than the overall task. The unconstrained dimensions listed in Table 2 were chosen in an effort to make the task as general as possible while keeping it easily reproducible. Random size, color, and orientation of the target, receptacle, and barriers make it difficult for simple heuristic solutions to succeed quickly. Limiting the vision of the human coach and the frequency of communication would place an increased problem-solving burden on the robot too. The task space represented by the SNR may be small, but it is relatively unstructured compared to many of the tasks robots currently perform.

3.6.3 SPECIFICITY

The numerical performance measure, p, described in Equation 1, gives the SNR specificity. It heavily rewards breadth, more so than excellence. The number of unlocked parameters, n, multiplicatively scales down the time required to do the task and scales up the total score. Performing a broad task slowly is likely to result in a higher p than performing a narrow task quickly.

3.6.4 LOW COST

The raw materials required to create the room and all its furnishings are relatively inexpensive and widely available (The LEGO Group, 2009). The bulk of the cost is in the CoroBot, which retails for less than $5,000.

3.6.5 Simplicity

The full description of the SNR given here occupies approximately one column and introduces no difficult concepts. At a more intuitive level, an image could convey a typical instance of the task with even greater economy, while maintaining a reasonable degree of accuracy. (See Figure 3.)

3.6.6 Range

By having a series of parameters that can be unlocked, the SNR can be performed by robotic systems with a wide variety of capabilities. With all free parameters locked, a carefully programmed robot could execute the task feed-forward, requiring no sensory feedback or intelligence at all. With all the parameters unlocked, the robot would be forced to navigate an unstructured environment with very little specific guidance from its coach. This would require continual adaptation to novel situations and pursuit of imprecisely defined goals. And should the SNR prove to be insufficiently challenging at some time in the future, it can easily be extended by adding additional parameters to Table 2. These might include multiple targets, greater variety in types of targets, moving elements, or competing robots.

3.6.7 Task Focus

By definition, performance on the SNR provides a task-focused benchmark. The measure, p, is a function only of the number of unlocked parameters and the time required to complete the task.

The SNR is an example of what an AGI battery task might look like. Generating an actual task would most likely be a collaborative, or at least iterative, process with other researchers providing suggestions, criticisms, and insights.

4. Discussion

This paper contains several novel contributions aimed at addressing the current lack of AGI benchmarks for natural world interaction. First, there is a list of criteria that may be used to evaluate benchmark candidates. The criteria are intended to ensure that an AGI benchmark would be both technically valid and practical to implement. In particular, they specify a task-based approach to AGI benchmarking.

Second, a long-term benchmark candidate is presented: the direction task. It meets the benchmark criteria well and shows promise as a useful artificial intelligence measure, particularly as AGI technologies mature. However, it is extremely broad and may be somewhat unwieldy for use in the near term.

The third proposal in this paper is for a near-term benchmark, the AGI battery. It consists of a collection of individual tasks. While each one is somewhat limited in breadth, collectively the set of tasks in the battery may be quite broad. This allows the battery to be somewhat more straightforward to evaluate and more appropriate to today's AGIs while still covering a large task space.

The final contribution of this paper is a description of a task that might be included in the AGI battery, the search and retrieve task (SNR). In it, an autonomous mobile

robot locates a target object and places it in a receptacle. The environment in which the task is completed is well-parameterized, but randomly defined. The SNR would require an AGI to demonstrate multiple abilities that are often considered critical to intelligent behavior, including navigation, understanding speech, and visual object recognition. More importantly, its inherent variability would prohibit heavy reliance on task modeling, but instead would require learning, adaptation, and problem solving during the course of the task.

4.1 The next steps

The SNR emphasizes certain capabilities, but does not touch on many others. As represented in Figure 2, any individual battery task covers only a small portion of the entire direction task. A full AGI battery would include a variety of tasks that is as wide as possible. Other battery tasks might be centered on natural language processing, physical interaction, problem solving, and all other capabilities considered to be part of natural world interaction. Ideally, all battery tasks would have a common interface protocol, a set of low-level inputs and outputs. This would allow an AGI candidate to be evaluated on multiple tasks without customization.

The direction task and the AGI battery are not intended to be final answers to the benchmarking problem. Rather they are presented in the spirit of the "straw man," an admittedly limited solution incarnated so as to invite criticism, suggestions for improvement, and counterproposals. There may be relevant criteria missing from the list in Table 1, better benchmarks than the direction task, and a better short-term solution than the AGI battery. It is hoped that these benchmarks 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

References

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

Asuncion, A., and Newman, D. 2007. UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html.

AUVSI. 2009. AUVSI Unmanned Systems Online. http://www.auvsi.org/competitions/water.cfm, Accessed September 22, 2009.

Bayer, S.; Damianos, L.; Hirschman, L.; and Strong, G. 2004. A Summary of Previous Grand Challenge Proposals for Cognitive Systems. Technical report, The

MITRE Corporation. Version 1.5, Prepared for DARPA IPTO, http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA458170 Accessed: February 25, 2010.

Brachman, R. J. 2006. (AA)AI more than the sum of its parts. *AI Magazine* 27(4):19–34.

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

Cohen, P. R. 2005. If not Turing's test, then what? *AI Magazine* 26(4):61–67.

CoroWare Inc. 2007. The CoroWare CoroBot. http://www.corobot.net/, Accessed September 22, 2009.

DARPA. 2007. DARPA Urban Challenge. http://www.darpa.mil/grandchallenge/index.asp, Accessed September 22, 2009.

Dillman, R. 2004. KA 1.10 Benchmarks for Robotics Research. Technical report, University of Karlsruhe. Sponsored: European Robotics Research Network.

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

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

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

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

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

Goertzel, B.; Arel, I.; and Scheutz, M. 2009. 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. http://www.agi-roadmap.org/images/HLAIR.pdf. Accessed September 21, 2009.

Griffin, G.; Holub, A.; and Perona, P. 2007. Caltech-256 Object Category Dataset. Technical Report 7694, California Institute of Technology. http://authors.library.caltech.edu/7694.

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

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

Kennedy, J. F. 1961. Man on the Moon Address. http://www.homeofheroes.com/presidents/speeches/kennedy_space.html, Accessed September 22, 2009.

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

Laird, J. E.; Wray III, R. E.; Marinier III, R. P.; and Langley, P. 2009. Claims and challenges in evaluating human-level intelligent systems. In *Proceedings of the 2009 Conference on Artificial General Intelligence*. Atlantis Press.

Lebiere, C.; Gonzales, C.; and Warwick, W. 2009. 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.

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

Michel, O.; Rohrer, F.; and van Bourquin, Y. 2008. Rat's Life: A cognitive robotics benchmark. In et al., H. B., ed., *Proc 2008 European Robotics Sympoiusm*, volume STAR 44, 223–232. Berlin Heidelberg: Springer-Verlag.

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

Moore, A. 1990. *Efficient Memory-Based Learning for Robot Control*. Ph.D. Dissertation, University of Cambridge.

Mueller, S. T., and Minnery, B. S. 2008. Adapting the Turing Test for embodied neurocognitive evaluation of biologically-inspired cognitive agents. In *Proc. 2008 AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*.

Netflix. 2009. Netflix Prize Homepage. http://www.netflixprize.com/, Accessed September 23, 2009.

Nilsson, N. J. 1995. Eye on the prize. *AI Magazine* 16(2):9–17.

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

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

The LEGO Group. 2009. The Official Web Site of LEGO(R) Products. http://www.lego.com/, Accessed September 22, 2009.

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

The RoboCup Federation. 2009b. RoboCupHome Homepage. http://www.ai.rug.nl/robocupathome/, Accessed November 12, 2009.

Tino, P.; Hammer, B.; and Bodén, M. 2007. *Perspectives of Neural-Symbolic Integration*, volume 77. Heidelberg, Germany: Springer-Verlag. chapter 5. Markovian bias of neural-based architectures with feedback connections, 95–133.

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

Wang, P. 2008a. Editorial: What makes JAGI special. *Journal of Artificial General Intelligence* 1:1–2.

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

Weng, J. 2009. Task muddiness, intelligence metrics, and the necessity of autonomous mental development. *Minds and Machines* 19:93–115.

Wray, R., and Lebiere, C. 2007. Metrics for cognitive architecture evaluation. In *Proceedings of the AAAI-07 Workshop on Evaluating Architectures for Intelligence*.