Intelligence Assessment for Early-Stage Software Systems Aimed at Human-Level, Roughly Human-Like AGI

Ben Goertzel
Novamente LLC

One of the many difficult issues arising in the course of research on human-level AGI is that of "evaluation and metrics" – i.e., AGI intelligence testing.

It's not so hard to tell when you've achieved human-level AGI -- though there is *some* subtlety here, which I'll discuss below. However, assessing the quality of incremental progress toward human-level AGI is a much subtler matter. In this essay I'll present some thoughts on this issue, culminating in a couple specific proposals:

- *Online School Tests*, in which AGIs are tested via their ability to succeed in existing online educational fora
- of more immediate interest, a series of tests called the *AGI Preschool Tests* (AIP Tests[1], for short), based on the notion of "multiple intelligences" (Gardner, 1983) and also on some novel ideas regarding learning-based intelligence testing.

The AIP Tests suggested here are specifically intended for AGI systems that control agents embodied in 3D worlds resembling the everyday human world, via either physical robots or virtually embodied agents. Very differently embodied AGI systems (e.g. systems to be initially taught purely via text without any simulated human-like or animal-like body) would potentially need qualitatively different testing methdologies.

Apologies and Warnings

Just to set expectations properly, I note up front that I am not going to articulate here any extremely crisp, simple AGI testing method that could easily be used to create some sort of "X Prize" analogue for AGI's. I have thought about the latter possibility a great deal and have come to the conclusion that it probably is not a good direction to follow. General intelligence, by its nature, is complex and multifaceted and doesn't lend itself to simple test scenarios; this is why human intelligence tests come in multiple flavors, all of which are long and contain numerous questions of various types. And, testing for particular sorts of intelligent accomplishments, while easier to do than testing for general intelligence, seems inevitably to lead down the road of encouraging test-focused narrow-AI development rather than AGI development focused on the creation of AGI systems with truly broad and creative intelligence. However (obviously, since I wrote this essay), I do think that AGI intelligence testing is an important area worth of thought and consideration in the AI field – even though the result of work in this area is

[1] Pronounced "ape tests" ;-)

likely to be a "AGI IQ tests" that, like the AIP tests suggested here, are complex and multifaceted rather than simple, crisp and elegant.

Another caveat is that I am not here attempting to approach the problem of assessing "general intelligence" in a truly mathematically broad sense, as addressed e.g. in Legg and Hutter's (2007) formalizations of the intelligence concept. This is the meaning of the qualifiers "Human Level, Human-Like" in the title of the essay. My goal is to explore ways of testing early-stage versions of AGI systems that are aimed at being as smart as humans (and potentially ultimately smarter), and at being smart in ways that are rougly similar to the ways humans are smart. Undoubtedly there are many other ways of being smart, including many that we humans would never recognize as intelligence.

Of course the notions of "human level" and "roughly human-like" are both vague, non-rigorous notions; and this doesn't mean they can't be use to conceptually motivate rigorous tests, but it does mean that human common sense is going to have to be applied to determine the contexts in which the tests proposed here should be applied. For example, an AGI system aimed solely at mathematical theorem-proving would fail most of the tests proposed here, even if it had incredibly general and deep intelligence in the mathematical domain; and this is not surprising because this AGI would manifestly fails to fulfill the "roughly human-like" criterion, even if it is intuitively "human-level" or even superhuman. Similarly an AGI successfully emulating chimp would fail most of the tests proposed here, even though it would constitute an extremely important and impressive achievement; and this is not surprising because this AGI would manifestly fail to fulfill the "human-level" criterion.

Why Not Focus on Testing Individual Components?

One suggestion that is sometimes made, regarding AGI testing, is: "If objectively assessing an AGI's overall functionality is such a complex matter, then why not just test its individual components and validate that they work really well? If an AGI has components that are better than the best-of-breed components in today's narrow-AI systems, this surely tells you something." My position is that this is a fatally flawed approach to assessing AGI intelligence. My focus here will be solely on testing holistic system functionality, not testing the functionality of individual components of AGI systems

Of course, systematic and comparative testing of system components can be valuable, and we have done plenty of it in the AGI projects I've been involved with: for instance, testing the MOSES procedure learning component we use in NCE/OpenCog against other program learning algorithms (see Looks, 2006; and testing the PLN probabilistic logic framework we use in NCE/OpenCog against other probabilistic logic approaches (see the Appendix of Goertzel et al, 2008). However, there is a very deep problem with this sort of testing as approach to AGI IQ assessment, which is that it is generally not meaningful to compare AGI system components against other AGI system components that look similar on the surface, but actually embody radically different theoretical assumptions, related to their different roles in the overall systems in which they are embedded. Apparently similar components may potentially play subtly different roles in the overall AGI systems in which they are embedded. To drive this point home

thoroughly, I will now spend a couple paragraphs recounting in moderate detail one example in which this problem arose in my own work, in the comparative testing of the PLN inference engine used in NCE/OpenCog with the NARS inference engine used in Pei Wang's NARS AGI system. (The details of this example may be slightly opaque to readers not familiar with NARS or PLN, but it seems hard to give a concrete, detailed example that would not be somewhat obscure in a similar way.)

An example of the kind of comparison we did[2] was the following sort of inference: consider

```
Ben is an author of a book on AGI <tv1>
This dude is an author of a book on AGI <tv2>
|-
This dude is Ben <tv3>
```

versus

```
Ben is odd <tv1>
This dude is odd <tv2>
|-
This dude is Ben <tv4>
```

Here each of the English statements is a shorthand for a logical relationship that in the AI systems in question is expressed in a formal structure; and the notations like <tv1> indicate uncertain truth values attached to logical relationships, In both NARS and PLN, uncertain truth values have multiple components, including a "strength" value that denotes a frequency, and other values denoting confidence measures. However, the semantics of the strength values in NARS and PLN are not identical.

Doing these two inferences in NARS you will get

```
tv3.strength = tv4.strength
```

whereas in PLN you will not, you will get

```
tv3.strength >> tv4.strength
```

The difference between the two inference results in the PLN case results from the fact that

```
P(author of book on AGI) << P(odd)
```

and the fact that PLN uses Bayes rule as part of its approach to these inferences.

My initial reaction, on getting these results, was that the NARS results seemed not to make intuitive sense, because I was sure that any intelligent human, on being presented with these inferences, would assign tv3.strength >> tv4.strength. However, when I discussed the issue with Pei Wang, the creator of NARS, he responded by saying,

---

[2] This comparative testing was done by Izabela Freire in 2002, using research funding provided by David Hart

roughly (I'm paraphrasing him loosely, with some risk of unintentional error) that there are other ways of indirectly accounting for the fact that

```
P(author of book on AGI) << P(odd)
```

in NARS, and thus that just feeding NARS the above syllogisms without other background knowledge is not a fair comparative test ... instead you'd need to compare NARS vs PLN on these syllogisms in the context of a rich database of background knowledge, with overall properties similar to those that one would find in a system that had gained its knowledge from life-experience.

This example illustrates the subtlety of comparatively testing inference engines (or AGI system components in general) from an AGI perspective. And it reinforces the notion that the right metrics for AGI systems will almost surely have to do with the overall behaviors of systems controlled by the AGI systems (for example embodied agents like physical or virtual robots), rather than concerning themselves with abstracted, lower-level functionalities like individual inference steps (which, even if they look very similar, may mean different things to different AGI systems or algorithms).

Testing different inference engines on the same formal structures, may not tell you much of anything if these different inference engines interpret these same formal structures differently. However, doing tests involving controlling robots or virtual agents, or holding English conversations, bypasses this problem via referring to an "objective" world whose interpretation is approximatively shared by the humans ultimately doing the evaluating.

Neither PLN nor NARS's inference engine is intended as a whole AGI system -- each one is intended as part of an overall AGI design, in which it receives outputs from certain other system components, and gives outputs to certain other system components. If the other components of NCE/OpenCog control PLN inputs/outputs in a manner that systematically differs from the way the other components of the NARS systems control NARS inference engine inputs/outputs, then this makes it very hard to compare the two inference systems. This is a subtler issue than it may at first seem, because the different manners of controlling inputs/outputs may embody different conceptual and semantic assumptions. It is logically quite possible that both PLN and NARS could work well within the systems they are designed for, but work poorly if swapped and placed into the contexts designed for each other – even if their inputs and outputs have the same syntactif form and closely related (but not identical) semantics.

Another, related, simpler point is that focusing on testing individual system components tends to lead AI developers down a path of refining system components for optimum functionality on isolated, easily-defined test problems that may not have much to do with general intelligence. It is possible of course that the right path to AGI is to craft excellent components (as verified on various isolated test problems) and then glue them together in the right way. On the other hand, if intelligence is in large part a systems phenomenon, that has to do with the interconnection of reasonably-intelligent components in a reasonably-intelligent way (as I have argued e.g. in Goertzel, 2006), then testing the intelligence of individual system components is largely beside the point: it may be better to have moderately-intelligent components hooked together in an AGI-appropriate way, than extremely-intelligent components that are not able to cooperate with other components sufficiently usefully.

Ultimately, studying the functionality of individual system components to assess overall system intelligence makes no more sense than studying the properties of a runner's muscles, heart, lungs etc. to assess how fast they can run. Of course, a runner's internal properties are going to be correlated with their speed, but these correlations are going to be complex and require much research to unravel, in part because of subtle dependencies between body parts. Whereas direct assessments of a runner's speed, or an AGI system's behaviors, are far less theory-laden and hence more appropriate as approximatively "objective" measures.

Online School Tests: A Pragmatic Replacement for the Turing Test

The classic approach to assessing whether an AI has achieved human-level general intelligence is the Turing Test (Turing, 1950), which measures the ability of an AI to fool humans, in a conversational context, into believing it's human.

However, the Turing Test has proved a singularly poor guide for the development of early-stage AGI systems. The Loebner Prize, which is given each year to the AI system that comes the closest to passing the Turing Test, has in practice had very little to do with real work toward general intelligence. An early-stage AGI is almost inevitably going to be far worse at holding humanlike English conversations than a well-crafted chatbot filled with a bunch of stock phrases but no real understanding.

On top of the "chatbot" problem, the Turing Test also has additional issues: it is obviously problematic for AGI approaches that are oriented toward making human-level and roughly *but not strictly* human-like AI systems. For an AI system like this, impersonating a human may not necessarily be a fair nor useful test. Is it really fair to demand that an AI be able to believably describe the feeling of a stomachache, a hangover, or warm rain falling lightly on the back of one's neck? This seems roughly as fair as demanding that a human be able to believably describe the particular psychological sensation of a hard drive failure, or the exquisite combination of joy and disturbance resulting from an overly rapid increase in the polygon resolution of one's fellow agents in a virtual world. Of course, Turing did not intend his test as a necessary criterion for human-level intelligence nor as a practical goal for AI development; in his original conception it was more of a challenge to those who conceive intelligence in non-functional terms.

But if we don't want to use the Turing test, what is the alternative for assessing achieved human-level, roughly human-like AGI? One approach, I suggest, is the "Online University Test." If an AI can get a BA degree at a real university, via online coursework only (assuming for simplicity courses where no voice interaction is needed, only textual and mouse-based communication), then I suggest we should consider that AI to have human-level intelligence. Note that the coursework spans multiple disciplines, and the details of the homework assignments and exams are not known in advance (otherwise students would be able to cheat too easily). Some basic social interaction and natural language communication are needed here, as well as understanding of course material, ability to do online research, and ability to solve problems. However, there is no requirement to be strictly humanlike in order to pass university classes.

There are also online high schools and even elementary schools[3], so one can also postulate related Online Highschool Test and Online Elementary School Tests -- though it is unclear how much easier these tests would be for AI systems, as for many AI systems, the hard parts will not be the course material itself, but rather the social and linguistic aspects of the online education (i.e., "figuring out what are the problems to be solved," rather than solving the problems). We may group all these possibilities under the heading of an "Online School Test" methodology.

The Online University Test is fine as a criterion for what it means to create a "human level, roughly humanlike AGI." But it isn't much use as a guide for incremental development towards this goal. Arguably, by the time one has a system that can pass the Online Elementary School Test, one has already passed the most difficult phases of AGI design and engineering. Thus, the really tricky question regarding evaluation and metrics regards how to measure the development of AI systems that haven't yet achieved the functionality of a human elementary school student. We may formulate this problem as the challenge of creating an appropriate series of Preschool AI Tests, with a goal of measuring an AI's incremental progress toward Elementary School Intelligence (ESI).

Challenges in Creating Preschool AI Tests

One of the major challenges in creating a Preschool AI Test is that different approaches to AGI may naturally be evaluated by different sorts of tests. Any set of tests one creates, with the view of measuring an AI system's incremental progress toward elementary-school intelligence, is naturally going to favor some paths to ESI over others. In spite of this inevitable bias, however, it seems important to articulate Preschool AI Tests anyway. If different approaches to AGI come along with different tests, this is not ideal, but is by no means an insuperable obstacle to progress. Competitive comparison of different approaches is one purpose of testing, but not the only one: well-crafted tests are also valuable simply for helping AGI developers to understand what their systems are capable of.

The specific Preschool AI Test approach I'll suggest in this essay is oriented toward AGI systems that are physically or virtuall embodied, and won't be directly applicable to other sorts of AGI systems. Some parts of the suggested approach will apply to (for instance) purely text-chat-based systems, others will not.

Another major challenge is the problem of "cheating." By this I don't mean cheating on the part of the AI (such as surreptitiously instant-messaging its creator for answers), but rather on the part of the AI designer. Over and over again, in the history of AI, we've seen the danger of "overfitting an AI system" to a specifically, narrowly defined goal or set of goals. Over and over again, it turns out that hacks or narrow-AI cleverness of various sorts can be used to achieve a set of specific goals which at first seemed to require general intelligence ... without really capturing the spirit in which the goals were originally proposed. One can substantially work around this problem by making one's test broad enough in nature, but this isn't as easy as one might think.

Due to these challenges, it seems to me that the most important assessments of intermediate stages of AGI development are necessarily going to be qualitative.

[3] e.g. http://www.e-tutor.com/elementary.php

Objectively measurable milestones are going to be very useful for testing, tuning and tweaking AGI systems -- when they are used in the context of a deep understanding and appreciation of the qualitative goals. But I suggest that Preschool AI Tests should not be used as the primary tool for structuring development of early-stage AGIs; rather, only as a tool for helping guide developers to maximize quantitative progress along lines that are qualitatively sensible in terms of a deep underlying cognitive/AI theory.

Naively, it might seem that creating a variety of different test problems (which is part of the approach I'm going to suggest later on in this essay) could circumvent the "cheating" challenge. However, a moment's consideration shows that diversity is not a cure-all. Suppose one poses 50 different test problems, qualitatively different in nature. One "trivial" approach to passing these tests would be to create a narrow-AI approach to each one of the 50 problems separately, and then wrap up these 50 specialized solutions inside a common external interface.

Furthermore, it's not wholly clear where the boundary between this trivial "cheating-based" approach and serious AGI design lies. For instance, suppose two of the 50 problems in one's test set involve navigation in complex environments. Is it "cheating" to create a specialized navigation process within one's AGI system, or not? Eric Baum (author of "What Is Thought?"; Baum, 2004) is one serious AGI thinker who believes that hard-wiring navigation into an AGI system is the correct thing to do: he strongly feels that the human brain has an in-built navigation module, and that an engineered AGI system should have one too. In my own work with the Novamente Cognition Engine and OpenCog Prime, we have implemented a hard-wired navigation system for practical applications, but for future development are leaning toward a middle path, in which certain high-level spatial-movement functions useful for navigation are exposed to the AI's learning algorithms primitives, but the AI system must learn to compose these functions into a real navigation algorithm. I think this can lead to a more flexible and adaptive navigation algorithm than directly providing the AI with navigation algorithms ... but, whether this difference would be apparent on a simple navigation-based test problems, is not clear. Quite possibly, hard-wired navigation algorithms could be humanly-tweaked to do very well on a couple narrow classes of navigation-based test problems, and a learning-based approach might have trouble competing. After all, not all humans are all that good at navigating, either. However, if an AI system had to learn to navigate in an unfamiliar sort of environment, then the learning-based approach would obviously be more powerful.

One obvious, partial solution to the cheating challenge is not to reveal to the AI nor the AI designer too many specifics of the tests, in advance. The general nature of the tests should be revealed, but not the details. For instance: Perhaps the testers could reveal that some tests will require moving around in crowded environments, but not the specifics of what sort of navigation testing will be done.

Another partial workaround is to test, not just what an AI system can do, but what it can learn based on certain types of feedback. For instance, one could test an AI's ability to navigate in a certain environment, then give it some lessons on navigation, and then see how well it is able to navigate after that. This is by no means an ironclad defense against cheating, because an AI designer could always program an AI with both the knowledge of navigation, and the propensity to pretend not to know how to navigate until navigation lessons have been received. One can work around this problem as well

to an extent, by using the Randomized Learning Based Test method that I'll describe below – but even this is not a complete solution. Ultimately, we are brought back to the point that qualitative assessment is going to be the most important thing at the AI preschool level. The purpose of tests and metrics, at this stage, is going to be to guide qualitative assessment, rather than to replace it.

Randomized Learning Based Testing Methodology

Let us define a Learning-Based Test as consisting of three parts:

1. a pre-test
2. some (generally interactive) instruction
3. a post-test, that measures how well the learner has learned from the instruction

For instance, one might

1. test an AI's ability to correctly identify the emotion associated with a gesture
2. give it some interactive instruction on identifying emotions associated with gestures (e.g. by explicitly telling it "When I do this I'm happy" while smiling; or else by giving it positive and negative reinforcement signals when it makes correct vs. incorrect judgments)
3. then re-test its ability

As noted above, the problem with this sort of test is that (to continue with the above example) an AI designer could potentially pre-program their AI system with the capability to associated emotions with gestures, and also with the propensity to feign ignorance about this until instructed. We may call this the "Nintendogs problem," as the popular virtual-pets game involves animated dogs that are preprogrammed to "act as if they're learning" various behaviors – when in fact the code for the behaviors is supplied in advance, along with code telling them to do these behaviors correctly only after receiving a certain amount of reinforcement.

A partial workaround for the Nintendogs problem is what I call Randomized Learning Based Testing Methodology, or RLB testing for short. RLB testing takes advantage of the fact that with AI's, unlike human children, it is possible to create multiple copies of the same AI and give each of them different instructions. However, it only works for the teaching of things that are in some sense arbitrary, rather than "natural." The idea is as follows:

1. Give an AI a pre-test
2. Then, create N copies of the AI, and place them out of communication with each other
3. Give each of the copies of the AI a separate instructional experience, aimed at teaching a somewhat different set of specific skills (but of the same general nature)
4. Give each of the copies of the AI a post-test, that measures how well the learner has learned from the instruction

So, for example, to continue with the emotion/gesture identification task, in the RLB method one of the copies might be taught that smiling indicates happiness, whereas another might be taught that it indicates anger, and another might be taught that it indicates sadness. The problem of course is that if the AI has been watching movies or studying images of people, it may have already learned that smiling really indicates happiness – so that some of the copies are being asked to learn plainly artificial, fake information, whereas others are being asked to learn information that is accurate in the context of the real world. On the other hand, any AI subjected to the test would be subjected to the same protocol, so there's nothing unfair about it.

Teaching an AI the rules of a game like baseball is another good example for this sort of methodology. The rules of baseball are fairly arbitrary, so that there should be no problem teaching different copies of an AI different variants of the rules. Furthermore, there are many different variants so it's not very likely that a clever, nefarious AI designer is going to preprogram their AI with a knowledge of all the variants that the clever, nefarious test designers are likely to cook up.

On the other hand, this sort of methodology seems less likely to be effective in contexts like language learning. Yet, even here there are some tests one could easily apply RLB too. For instance, one could make up fake words of different types – one could teach copy 1 of the AI the proper use of "fnorbulate", teach copy 2 of the AI the proper use of the word "gttrbuckular", and so forth. This would certainly test the ability of the AI to learn usage of different words of different sorts. Making up new grammatical rules to teach different copies is harder because we don't know as much about what makes a "psychologically natural" grammar rule, and it's harder for us as teachers to effectively and naturalistically use a made-up grammar rule, as opposed to a made-up word.

RLB is not a sufficiently powerful idea to fully overcome the cheating challenge associated with AGI testing, but, it does seem a worthwhile addition to the arsenal of AGI testing methodologies.

The Multiple Intelligences Approach

The specific approach I suggest for an AGI Preschool Test, for the case of AGI systems with roughly humanlike physical or virtual embodiment, is based on the learning based and RLB testing methodologies introduced above, combined with the psychological notion of multiple intelligences.

"Multiple intelligences" is a psychological approach to intelligence assessment based on the idea that different people have mental strengths in different high-level domains, so that intelligence testing should contain tests that focus on each of these domains separately. My suggested use of the multiple intelligences framework for AGI is not particularly tied to the value (or otherwise) of the framework for assessing human intelligence. The value of the framework for assessing AGI intelligence lies in its explicit attention to the broad, general scope of human intelligence.

The following table[4] summarizes the key intelligences posited within the theory:

Intelligence	Aspects	Tests
Linguistic	words and language, written and spoken; retention, interpretation and explanation of ideas and information via language, understands relationship between communication and meaning	write a set of instructions; speak on a subject; edit a written piece or work; write a speech; commentate on an event; apply positive or negative 'spin' to a story
Logical-Mathematical	logical thinking, detecting patterns, scientific reasoning and deduction; analyse problems, perform mathematical calculations, understands relationship between cause and effect towards a tangible outcome or result	perform a mental arithmetic calculation; create a process to measure something difficult; analyse how a machine works; create a process; devise a strategy to achieve an aim; assess the value of a business or a proposition
Musical	musical ability, awareness, appreciation and use of sound; recognition of tonal and rhythmic patterns, understands relationship between sound and feeling	perform a musical piece; sing a song; review a musical work; coach someone to play a musical instrument; specify mood music for telephone systems and receptions
Bodily-Kinesthetic	body movement control, manual dexterity, physical agility and balance; eye and body coordination	juggle; demonstrate a sports technique; flip a beer-mat; create a mime to explain something; toss a pancake; fly a kite; coach workplace posture, assess work-station ergonomics
Spatial-Visual	visual and spatial perception; interpretation and creation of visual images; pictorial imagination and expression; understands relationship between images and meanings, and between space and effect	design a costume; interpret a painting; create a room layout; create a corporate logo; design a building; pack a suitcase or the boot of a car
Interpersonal	perception of other people's feelings; ability to relate to others; interpretation of behaviour and communications; understands the relationships between people and their situations, including other people	interpret moods from facial expressions; demonstrate feelings through body language; affect the feelings of others in a planned way; coach or counsel another person

[4] This table is borrowed with minor modifications from www.businessballs.com/howardgardnermultipleintelligences.htm

Whether all the intelligences in this table are necessary to consider from an AGI perspective is not clear. The necessity of the linguistic, interpersonal, spatio-visual and logico-mathematic intelligences is obvious: without these, there is no way an AI will pass the Online Elementary School Test, for example. Musical intelligence can potentially be ignored for the purpose preparing an AGI for Online School Tests; but the situation with Bodily-Kinesthetic intelligence is less clear: it may be that achieving some measure of Bodily-Kinesthetic intelligence is going to be critical for the understanding of linguistic metaphors related to bodily-kinesthetic activity, which are rampant in ordinary language.

My specific suggestion for testing preschool-level AGI systems is to create a number of test categories based on each of the multiple intelligences listed above (and the phrases in the third column are examples of potential test categories). Then, for each of these categories, multiple specific tests may be generated, using learning-based and RLB testing whenever possible. To have a good testing methodology, AGI's and their developers shouldn't know the specific tests to be used in advance anyway, but only the general categories. Specific examples of tests within each category should be provided for guidance, but the actual tests given should not rigidly imitate the specifics of the example tests.

I will here give examples of possible tests within each of the five types of intelligence mentioned above (excluding only musical). In the examples I'll use the case of an AI controlling agents in virtual worlds, for sake of concreteness, but the same examples obviously apply to physical robotics.

A Linguistic Test

An example test of linguistic intelligence is the task of writing a set of instructions. Suppose we have two human-controlled avatars, A and B, and one AI-controlled avatar. And, suppose A shows the AGI how to carry out some task X, and then leaves. The AI's job is then to show B how to do that same task X.

This has many variants, including cases where the best way to describe X is purely verbal, and others where the best way to desribe X involves a combination of words and actions.

A concrete example would be teaching someone how to assemble a piece of furniture, similar to the furniture kits one buys at K-mart or Staples ... or a bicycle. Of course, the specific type of item to be assembled would not be known to the AGI or AGI designer prior to the test being given.

Using the RLB methodology, the AGI's could be given a period of feedback regarding how well they gave instructions ... and then after the feedback period, could be tested on how well they absorbed the instructions.

A Logico-Mathematical Test

Example tests of logico-mathematical intelligence are the task of creating a process to measure something difficult, or compare two or more entities regarding some quantity that is difficult to measure. One paradigm that can be used here is that of indirect comparisons. To quote extensively from (Reece et al, 2001)

Indirect comparisons require the ability to make two kinds of mental relationships - transitive reasoning and unit iteration - which are best explained by the following task. Piaget et al. asked children in individual interviews to build a tower having the same height as a model 80 cm tall (Piaget et al., 1948/1960). The child's tower was to be built on a table 90 cm lower than the base of the model. The child was given smaller blocks than the ones used in the model so that one-to-- one correspondence was impossible. Long strips of paper, as well as a ruler and three sticks were provided - a stick 80 cm long, one that was longer, and one that was shorter than 80 cm.

Before the age of about seven, children did not use the sticks or ruler; and when 3-, 4-, and 5-year-olds were asked if a stick or a ruler might be useful, they answered "No." Five-year-olds consistently wanted to bring the two towers together for direct comparison, but the interviewer did not allow this action. Children then used various body parts in an attempt to compare the two towers as precisely as possible.

Finally, around the age of seven, children began to use one of the longer sticks as a third term. (Here, the model tower and the copy were the first two terms, and the stick was the third.) Seven-year-olds marked the height of the model on the stick, took the stick to their tower, and made their tower as tall as the height indicated on the stick. Piaget et al. explained this use of a stick as a manifestation of transitive reasoning that becomes possible when the child's logic has developed.

Transitivity refers to the ability to deduce a third relationship from two (or more) other relationships of equality or inequality. The child who can reason transitively can deduce that if the height of the model tower and the length marked on the stick are equal (a direct comparison), and this length marked on the stick is equal to the height of his or her tower (another direct comparison), then the height of the two towers must be the same (a deduction). Most children before the age of seven cannot understand this logical reasoning, even if it is explained to them.

Piaget et al. went on to show a small block to the children and asked if it could be used to compare the height of the two towers. The children who had demonstrated transitive reasoning responded in one of two ways: The less advanced group said that the small block was too small to be of any use, but the more advanced group used it as a unit to iterate and count. These children placed the small block at the bottom of the tower, marked the upper end of the block on the tower, moved the block up until its lower end was exactly on the mark, and repeated the same procedure upward to the end of the tower, without any gaps or overlappings. These actions showed that the child thought

about the length of the block as a part of the tower's height. Recent research indicates that a majority of children become able to iterate a unit of length around third grade (Kamii & Clark, 1997). Note that when children have developed the logic of unit iteration, their measurement becomes exact.

It is easy to see how one could create a variety of indirect comparison tests along the lines of the above. Furthermore, one could easily use an RLB approach in this case, giving different sorts of structures and measuring implements to different replicates of an AI, to avoid the risk of the AI being specialized to some particular type of structure or measuring implement.

A Bodily-Kinesthetic Test

An example bodily-kinesthetic test would be the ability to communicate observed activities using mime. The AGI would watch human agent A carry out a certain action involving other agents or objects; then these agents or objects would be removed from the scene, and the AGI would need to do a mime for human agent B, indicating to A the activity in question. This is basically the game of "charades."

Another example would be the ability to teach another agent a dance. Human agent A teaches the AGI a dance, and then goes away; and the AI is then supposed to teach human agent B the dance. The AGI will have to demonstrate the dance, but then also correct B if B does it wrong, and explain the right way to do it. This particular example would be hard to do given current virtual world technology, but would be easy in near-future virtual worlds enabled with better haptic devices and finer-grained avatar control.

A Spatial-Visual Test

An example test of visuospatial intelligence is the creation of a room layout. Suppose an AGI is told what people are going to live in a house, and a few things about them, including what their tastes and occupations are. Then the AGI has to figure out what furniture they need and how to arrange it. The human occupants then rate the room layout based on how much they would like to live in the room. This lends itself well to RLB since different people may have very different tastes.

Another test would be the ability to draw "cave paintings" – i.e., given a simple marker or paintbrush, to create images that evoke particular objects. The AGI would be shown an object, and would then need to draw a picture conveying the object to a human viewer, the accuracy being judged by whether the human could correctly identify the object (from among a long list of choices). To make the test more interesting, using an RLB approach one could have the final test involve a variety of different artistic media: pens, paint ... rocks arranged on the ground, etc.

An Interpersonal Test

An example test of interpersonal intelligence is the recognition of feelings through body language or tone of voice. Recognition of feelings is an interesting test task

because of how well it fits in with the RLB methodology: one can easily have human testers express their feelings in odd ways (different ways to each AGI copy) and see how well the AGI adapts. Also, one can use testers from different cultures, who habitually express feelings in different ways. The measurement of accuracy is easy here of course: one simply asks the AGI what the human it's interacting with is feeling.

Virtual worlds are fairly weak for expression of feeling except through voice, but use of haptic interfaces and cutting-edge virtual-world technology would make this sort of testing possible. One could also simply use videos of human faces for this sort of task, though this requires AGIs with strong vision processing components.

Another interpersonal intelligence test, not requiring so much on the perception side, is listening to a conversation between people (preferably in an embodied context where the AGI can see the world the people are talking about) and telling when they are joking. Again accuracy assessment is pretty easy here: one just needs the AGI to report when it thinks the people are joking. This lends itself very well to an RLB approach because different people can have such different senses of humor: the "humor teachers" may well have different senses of humor than the conversors that the AI listens to during the final testing phase.

Conclusion

While I have not specified a concrete, usable "AGI IQ test" here, I believe I have laid out a direction along which such a test could practicably be constructed. The next step would be to make the ideas of the prior section more concrete, and create a variety of conceptually similar tests embodying different test categories probing the multiple intelligences.

Ideally, one would like to see a number of different researchers, proponents of different designs aimed at human-level, roughly human-like AGI, agree to a common testing approach, such as a specific incarnations of the Online School Test and AGI Preschool Test proposed above. Such agreement could be a valuable step in terms of crispening the focus of the AGI research community.

References

- Baum, Eric (2004). What is Thought?, MIT Press
- Gardner, Howard. (1983) "Frames of Mind: The Theory of Multiple Intelligences." New York: Basic Books.
- Goertzel, Ben (2006). The Hidden Pattern. Brown Walker Press.
- Goertzel, Ben, Matthew Ikle', Izabela Freire Goertzel and Ari Heljakka (2008). Probabilistic Logic Networks. Springer.
- Legg, Shane and Marcus Hutter (2007). Universal Intelligence: A Definition of Machine Intelligence. In Minds and Machines, pages 391-444, volume 17, number 4, November 2007.
- Looks, Moshe (2006). Competent Program Evolution. PhD Thesis in Computer Science Department, Washington University, St. Louis.
- Reece, Charlotte Strange, Kamii, Constance (2001). The measurement of volume: Why do young children measure inaccurately?, School Science and

Mathematics, Nov 2001, http://findarticles.com/p/articles/mi_qa3667/is_200111/ai_n9009076

- Turing, Alan (October 1950), "Computing Machinery and Intelligence", Mind LIX(236): 433–460, http://loebner.net/Prizef/TuringArticle.html
- Wang, Pei (2006). Rigid Flexibility: The Logic of Intelligence. Springer.