# Toward a Roadmap for Human-Level Artificial General Intelligence: Embedding HLAI Systems in Broad, Approachable, Physical or Virtual Contexts

Preliminary Draft

Ben Goertzel and Itamar Arel and Matthias Scheutz

April 18, 2009

**Abstract**

We present the first steps of the creation of a roadmap toward human-level artificial general intelligence (HLAI). The intended roadmap is not targeted at any specific HLAI design or architecture, but will enable the development of multiple HLAI systems along the same pathway, toward the same end goals and passing through the same set of milestones. To that end, we assume that (1) 3D virtual or robotic embodiment should be utilized, and (2) the context with which an HLAI system is presented should possess two properties we label *breadth* and *approachability*.

## 1 Introduction

The purpose of this document is to propose an *approach to creating a roadmap* – not the roadmap itself – for the development of human-level, roughly "human-like", artificial general intelligence (HLAI, for short). Thus human-like AGI, rather than *any* AGI, is at the center of attention. The approach is based on core assumptions made about what paths could lead to HLAI. By labeling *context* as an environment coupled with a set of goals that are defined relative to that environment, the core assumptions pertain to three properties of the context given by the testing/training environment and the task: (1) *breadth* (i.e., both richness and diversity), (2) *accessibility* to systems without any specialized knowledge beyond that which is implicit in human-like intelligence, and (3) the *use of a complex 3D physical or virtual embodiment*. These assumptions are intended to be particular enough to facilitate the creation of a specific roadmap, yet not too particular so as to imply commitment to any underlying AGI architecture or technology.

The components that are perceived to be inherent to a HLAI roadmap are described, and some remarks are made about the contents that should comprise these components based on the core assumptions. The crux of the roadmap proposed would be a sequence of "milestone" tasks, occurring in a small set of common environments, organized so as to lead to a commonly agreed upon set of long-term goals.

Some of the ideas presented here arose during two workshops on "*Evaluation and Metrics for Human-Level AI*" organized by John Laird and Pat Langley (one in Ann Arbor in late 2008, and one in Tempe in early 2009). Some of the conclusions of the Ann Arbor workshop were recorded in (Laird et al, 2009). Inspiration was also obtained from discussion at the post-conference workshop of AGI-09, triggered by Itamar Arel's presentation on the "AGI Roadmap" theme (Arel, 2009a, Arel 2009b).

## 2 Assumptions

At the Tempe workshop mentioned above, participants identified two dimensions along which one might categorize contexts for training, teaching, evaluating and testing early-stage HLAI systems. Since there was little agreement on the best terminology for describing these dimensions, the following terms for describing two critical dimensions are introduced and used throughout the document: *breadth* (e.g., broad vs. narrow) and *accessibility* (e.g., accessible vs. knowledge-intensive). These terms are discussed here as properties of *contexts* for HLAI systems.

### 2.1 Breadth

The term "broad" is intended to capture the richness and diversity of everyday human reality. Such richness is typically reflected by a stochastic, dynamic, partially observable multi-agent environment which provides high-dimensional stimuli with different scales of spatiotemporal dependencies. It also includes the great variety of different types of tasks that humans perform on a daily basis.[1] As a result, any AGI context (i.e., environment plus task) that does not reach the complexity of everyday human reality is impoverished, but there are obviously wide variations in the degree of impoverishment.

There are several reasons why broad context is important for defining and evaluating HLAI systems. For one, broad context is what humans face and cope with. Hence, HLAI systems need to demonstrate competence and appropriate performance in new tasks they confront (possibly via drawing subtle analogies to other tasks they've encountered, even though these other tasks may seem quite different on the surface). Moreover, to be comparable to human behavior and performance, HLAI systems should approach most tasks in a way that has as much to do with the broader environmental and life context in which they experience the task, as with the abstract structure of the task. Finally, HLAI systems should be "open" in the sense of being able to *be confrontable* with new tasks in new environmental settings without simply failing at them (as would be the case with most if not all current systems). This is last capability of HLAI systems is effectively that of "universal learners": for a given task $T$ within its environment $E$, a *universal learner* can learn how to perform it, but it might not be as efficient as a special purpose system that has been optimized to do $T$ and only $T$. The trade-off comes at the architectural level (assuming that the physical sensor and effector requirement for $T$ are met), including the ability to represent the task in a way that allows for the discovery of rules and control principles to achieve it.

### 2.2 Accessibility

The term "accessible" is intended to capture properties of an environment/task context which allow an HLAI system with minimal pre-programmed, built-in knowledge (e.g., the equivalent of the genetically encoded knowledge in humans) to succeed at its tasks. Like breadth, this is somewhat of a fuzzily-defined quantity, since it in some sense depends on aspects of human nature which are not yet fully understood. The boundary between knowledge acquired through experience and that which is genetically given in humans is subtle and still being studied. Nonetheless, it is hoped that the intention of the "accessibility" assumption will be qualitatively clear.

[1] Notice that when we contrast *broad* with *narrow*, we do not intend the reading associated with "narrow AI" . The game of Go, for example, is a " broad" context in some sense as it provides a wide variety of different complex mathematical and interactional patterns – yet, it exhibits only a tiny fraction of the phenomena that characterize the typical human environments. Thus, from an everyday human perspective, it does not satisfy the requirements of broad context.

In contrasting "accessible" with " knowledge-intensive", one isolates scenarios in which a massive advantage is given to AGI systems that are pre-programmed with specialized knowledge. Examples of the latter would be a knowledge-intensive game like FreeCiv, or a domain like medical question answering.[2] One strong reason to prefer accessible contexts would if one suspects there are severe limitations on the degree to which an AGI system can flexibly generalize a piece of knowledge X, when it has not acquired X through its own experience.

### 2.3 3D Embodiment

Given the general assumptions of breadth and accessibility as defined above, a number of possibilities exist regarding environments for AGI development, including

1. embody the AI in a chatbot with Internet access (here breadth is arguably provided via the rich variety of multimedia and social media available online)
2. embody the AI in a 3D or 3D virtual world (not necessarily as autonomous agent)
3. embody the AI in a robot

While all of the above have some viability to them, in order to progress toward a concrete roadmap, a choice must be made; and for the proposed roadmapping effort options 2 and 3 are selected. However, the current perception is that it will be possible to pursue a roadmap encompassing 2 and 3 as coexisting alternatives. One can articulate closely parallel milestones in virtual and physical environments, as long as the virtual environments . Of course, some milestones may be far easier to achieve in a virtual world than in the physical world, but that is an issue pertaining to development timelines rather than the main body of the roadmap.

### 2.4 Reality Testing

It is further believed that it is important for the roadmapping process to include contexts in which humans either regularly do or at least could participate. This gives one a very robust method of realistically assessing the breadth and accessibility of the context. Without this "human touchstone" , it may be all too easy to fall prey to the various illusions and pitfalls commonly associated with "toy environments" in historical AI research. Thus, ideally, physical or virtual environments utilized for HLAI systems should be ones that humans can spend a lot of time in, where the humans interact with the environment in roughly the same way that the HLAI systems do.

## 3 Overview of the AGI Roadmapping Process

A variety of approaches to technology roadmapping exist; a roadmapping schematic involving the following steps is assumed (which have been customized for the specific case of HLAI):

1. identify the "end product" that will constitute the focus of the roadmap

[2] While we cannot rule out contexts in which AGI systems with preprogrammed knowledge (e.g., of fundamental notions like time, space and agency) will have significant advantage, we want to avoid contexts that give significant advantage to AGI systems just because they have some specialized pre-programmed knowledge that goes beyond that which human brains can be plausibly hypothesized to contain via their genetic endowment. As such, preprogrammed knowledge should be viewed an added benefit (often dramatically so) rather than a must for any AGI system.

2. identify the environments and tools that will be used to enable the following of the roadmap
3. Specify major "incremental capabilities" or "milestones" that build toward the end product
4. Specify methods of evaluating and measuring these incremental capabilities
5. identify technology alternatives and their projected timelines

Our goal here will not be to explore any of these steps exhaustively, but merely to give a rough indication of how each aspect might be fleshed out in the course of a roadmapping effort. Point 2 "environments" was already addressed above as part of discussing our assumption about 3D physical or virtual environments (which should live up to the challenges of the real-world, i.e., real-time, real metric space, real physics, etc.). Next, the other steps are addressed in sequence.

### 3.1 Long-Term Goals of HLAI

What is the outcome or product at the end of the proposed roadmap? The end goal of work toward HLAI is a question of significant subtlety. Examples of potential end goals that HLAI researchers find relevant are:

- pass the *Turing Test*, conceived as (something like) "fooling a panel of college-educated human judges, during a one hour long conversation, that one is a human being"
- pass the *Total Turing Test*, which requires the system in a robotic instantiation (i.e., in a face-to-face conversation)
- pass the *Virtual World Turing Test* occurring in an online virtual world, where the HLAI and the human controls are controlling avatars (this is inclusive of the standard Turing Test if one assumes the avatars can use language)
- pass the *Telerobotic Turing Test*, which is similar to the virtual world Turing Test, but the HLAI and the human controls are remote-controlling robots in some test environment (which would be best chosen as a real-world outdoor environment; this one is also inclusive of the Turing Test)
- pass a *general IQ test*, *test of G*, *Mensa membership* or any other comprehensive standardized test
- pass the *Online University Test*, where an HLAI has to obtain a college degree at an online university, carrying out the same communications with the professors and the other students as a human student would (including choosing its curriculum, etc.)
- pass the *Artificial Scientist Test*, where an HLAI that can do high-quality, original scientific research, including choosing the research problem, reading the relevant literature, writing and publishing the paper, etc. (this may be refined to a *Nobel Prize Test*, where the HLAI has do original scientific research that wins a Nobel Prize)

It's worth emphasizing that the above are not suggested as end goals of AGI research in the sense that it is believed they can never be superseded. There is the possibility that in the future AGI may advance beyond the human level; but our goal here is to discuss a roadmap toward human-level, roughly human-like AGI in particular.

Another point to stress is that most of the above end goals have a human-centric theme to them, in that humans are inherently involved in the context considered. That is definitely true for the Turing Test. The idea that a human-centric testbed is of critical value and should, particularly toward the end of the roadmap, play a key role, is broadly embraced.

### 3.2 Specifying Incremental Capabilities: Taking Human Cognitive Development as a Guide

Based on the assumptions articulated above, there seems to be a very natural approach to creating a set of incremental capabilities building toward HLAI: *to draw on our copious knowledge about human cognitive development*. This is by no means the only possible path; one can envision alternatives that have nothing to do with human development (and those might also be better suited to non-human AGIs). However, so much detailed knowledge about human development is available – as well as solid knowledge that the human developmental trajectory does lead to human-level AI – that the motivation to draw on human cognitive development is quite strong.

The main problem with the human development inspired approach is that cognitive developmental psychology is not as systematic as it would need to be for HLAI to be able to translate it directly into architectural principles and requirements. While early thinkers like Piaget and Vygotsky outlined systematic theories of child cognitive development, which are no longer considered accurate, one currently faces a mass of detailed theories of various aspects of cognitive development, but without an unified understanding. Nevertheless it is believed that it will be viable to work from the human-development data and understanding currently available, and craft a workable HLAI roadmap therefrom.

### 3.3 Evaluating Incremental Progress

Regarding specific tests and metrics for system evaluation, the choice of incremental capabilities founded on human developmental psychology has the advantage of allowing HLAI research to draw on a vast body of knowledge from developmental and educational psychology regarding evaluation and measurement of child cognitive ability. For each chosen capability, one may inspect the experimental child cognitive developmental psychology literature to see what experimental protocols are used to measure that capability, and then adapt these protocols to the case of (robotic or virtual) HLAI systems.

The implementation of some such tasks and tests in a "virtual school" context is discussed in (Goertzel and Bugaj, 2009), though this is far from the only viable approach. [3]

### 3.4 Identifying Technology Alternatives and their Timelines

Once the above items are agreed upon, one may then invite HLAI system designers to create projections regarding the timelines according to which their systems might be able to achieve the steps identified, with various degrees of effectiveness. The quantitative timelines obtained here will be interesting to study; but at least as interesting will be the relative difficulty levels assigned to different capabilities by different designers, and the temporal orderings of capabilities that different HLAI designers find most natural according to their designs. As most designs aimed at HLAI do not attempt to closely emulate the human mind or brain, there is no reason to expect that they will necessarily assign the same relative difficulty levels or temporal ordering to the capability

[3] A simple example maybe be the "brainquest" cards which are intended to test the cognitive abilities of different ages ranging from 3 to 15, see brainquest.com.

milestones. Differences in ordering or difficulty ranking may be reflective of differences of underlying design principles.

# 4 Conclusions

We next outline the basics of what are believe to be a workable path toward creating a HLAI roadmap that would be useful to the HLAI research community. In order to achieve the concreteness necessary for such a pursuit, it is posited that (1) HLAI research programs focus on broad, accessible contexts, and that (2) milestones be drawn from human cognitive development.

Some comments on each of the five steps in the above roadmapping overview follow:

1. Agreeing on a set of tests as described in Step 1 may require some discussion, while refinement of the tests agreed upon in Step 1 may require some work, but none of this seems likely to be extremely problematic.

2. Step 2 is relatively simple in concept. There may be a bifurcation between researchers who are committed to physical robotics and those who are open to virtual worlds, but aside from this no deep issues seem likely to arise. There may be differences of opinion regarding the degree of breadth that is important to provide, or the nature of accessibility.

3. It would appear that the most subtantial step in creating the roadmap is Step 3: the identification of, and agreement on, a set of capabilities, demonstrable in the environment(s) described in Step 2, and leading toward HLAI as defined in Step 1.

4. Once Step 3 is complete, Step 4 appears to be a matter of systematic work and detail-refinement, on which different researchers are unlikely to disagree profoundly.

5. Step 5 is then work which different HLAI designers would do independently on their own – perhaps they will not agree with each others' timelines, but at least at this stage there will be a clear collective understanding of what is being discussed.

6. Our practical suggestion is to gather a group of HLAI researchers interested in pursuing a roadmapping exercise as described above (based on the presented assumptions), and for this group to collectively and systematically proceed through steps 1-5 as defined above.

**References**


- Arel, Itamar (2009). Working Toward Pragmatic Convergence: AGI Axioms and a Unified Roadmap. AGI-09 Workshop on the Future of AGI

- Arel, Itamar and Scott Livingston (2009). Beyond the Turing Test. IEEE Computer, Vol. 42, No. 3, pp. 90-91, March 2009

- Goertzel, Ben and Stephan Vladimir Bugaj (2009). AGI Preschool. Proceedings of the Second Conference on Artificial General Intelligence (AGI-09), Atlantis Press

- Laird, John, Robert Wray, Robert Marinier and Pat Langley (2009). Claims and Challenges in Evaluating Human-Level Intelligent Systems. Proceedings of the Second Conference on Artificial General Intelligence (AGI-09), Atlantis Press