

Artificial General Intelligence: Concept, State of the Art, and Future Prospects

Ben Goertzel

BEN@GOERTZEL.ORG

OpenCog Foundation

G/F, 51C Lung Mei Village

Tai Po, N.T., Hong Kong

Editor: Tsvi Achler

Abstract

In recent years broad community of researchers has emerged, focusing on the original ambitious goals of the AI field – the creation and study of software or hardware systems with general intelligence comparable to, and ultimately perhaps greater than, that of human beings. This paper surveys this diverse community and its progress. Approaches to defining the concept of Artificial General Intelligence (AGI) are reviewed including mathematical formalisms, engineering, and biology inspired perspectives. The spectrum of designs for AGI systems includes systems with symbolic, emergentist, hybrid and universalist characteristics. Metrics for general intelligence are evaluated, with a conclusion that, although metrics for assessing the achievement of human-level AGI may be relatively straightforward (e.g. the Turing Test, or a robot that can graduate from elementary school or university), metrics for assessing partial progress remain more controversial and problematic.

Keywords: AGI, general intelligence, cognitive science

1. Introduction

How can we best conceptualize and approach the original problem regarding which the AI field was founded: the creation of thinking machines with general intelligence comparable to, or greater than, that of human beings? The standard approach of the AI discipline (Russell and Norvig, 2010), as it has evolved in the 6 decades since the field's founding, views artificial intelligence largely in terms of the pursuit of discrete capabilities or specific practical tasks. But while this approach has yielded many interesting technologies and theoretical results, it has proved relatively unsuccessful in terms of the original central goals of the field.

Ray Kurzweil (Kurzweil, 2005) has used the term “narrow AI” to refer to the creation of systems that carry out specific “intelligent” behaviors in specific contexts. For a narrow AI system, if one changes the context or the behavior specification even a little bit, some level of human reprogramming or reconfiguration is generally necessary to enable the system to retain its level of intelligence. This is quite different from natural generally intelligent systems like humans, which have a broad capability to self-adapt to changes in their goals or circumstances, performing “transfer learning” (Taylor, Kuhlmann, and Stone, 2008) to generalize knowledge from one goal or context to others. The concept of “Artificial General Intelligence” has emerged as an antonym to “narrow AI”,



to refer to systems with this sort of broad generalization capability.^{1 2} The AGI approach takes “general intelligence” as a fundamentally distinct property from task or problem specific capability, and focuses directly on understanding this property and creating systems that display.

A system need not possess infinite generality, adaptability and flexibility to count as “AGI”. Informally, AGI may be thought of as aimed at bridging the gap between current AI programs, which are narrow in scope, and the types of AGI systems commonly seen in fiction – robots like R2D2, C3PO, HAL 9000, Wall-E and so forth; but also general intelligences taking non-robotic form, such as the generally intelligent chat-bots depicted in numerous science fiction novels and films. And some researchers construe AGI much more broadly than even the common science fictional interpretations of AI would suggest, interpreting it to encompass the full gamut of possible synthetic minds, including hypothetical ones far beyond human comprehension, such as uncomputable minds like AIXI (Hutter, 2005). The precise definition or characterization of AGI is one of the subjects of study of the AGI research field.

In recent years, a somewhat broad community of researchers united by the explicit pursuit of AGI has emerged, as evidenced for instance by conference series like AGI³, BICA⁴ (Biologically Inspired Cognitive Architectures) and Advances in Cognitive Systems⁵, and numerous special tracks and symposia on Human-Level Intelligence⁶, Integrated Intelligence⁷ and related themes. The “AGI community”, consisting e.g. of the attendees at the AGI-related conferences mentioned above, is a fuzzy set containing researchers with various interpretations of, and varying levels of commitment to, the AGI concept. This paper surveys the key ideas and directions of the contemporary AGI community.

1.1 What is General Intelligence?

But what is this “general intelligence” of what we speak? A little later, I will review some of the key lines of thinking regarding the precise definition of the GI concept. Qualitatively speaking, though, there is broad agreement in the AGI community on some key features of general intelligence:

- General intelligence involves the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments.
- A generally intelligent system should be able to handle problems and situations quite different from those anticipated by its creators.

1. Kurzweil originally contrasted narrow AI with “strong AI”, but the latter term already has a different established meaning in the AI and cognitive science literature (Searle, 1980), making this an awkward usage.

2. The brief history of the term “Artificial General Intelligence” is as follows. In 2002, Cassio Pennachin and I were editing a book on approaches to powerful AI, with broad capabilities at the human level and beyond, and we were struggling for a title. I emailed a number of colleagues asking for suggestions. My former colleague Shane Legg came up with “Artificial General Intelligence,” which Cassio and I liked, and adopted for the title of our edited book (Goertzel and Pennachin, 2007). The term began to spread further when it was used in the context of the AGI conference series. A few years later, someone brought to my attention that a researcher named Mark Gubrud had used the term in a 1997 article on the future of technology and associated risks (Gubrud, 1997). If you know of earlier published uses, please let me know.

3. <http://agi-conf.org>

4. <http://bicasociety.org>

5. <http://www.cogsys.org/>

6. <http://www.aaai.org/Press/Reports/Symposia/Fall/fs-04-01.php>, http://www.ntu.edu.sg/home/epnsugan/index_files/SSCI2013/CIHLI2013.htm

7. <http://www.aaai.org/Conferences/AAAI/2011/aaai11icall.php>

- A generally intelligent system should be good at generalizing the knowledge it’s gained, so as to transfer this knowledge from one problem or context to others.
- Arbitrarily general intelligence is not possible given realistic resource constraints.
- Real-world systems may display varying degrees of limited generality, but are inevitably going to be a lot more efficient at learning some sorts of things than others; and for any given real-world system, there will be some learning tasks on which it is unacceptably slow. So real-world general intelligences are inevitably somewhat biased toward certain sorts of goals and environments.
- Humans display a higher level of general intelligence than existing AI programs do, and apparently also a higher level than other animals.⁸
- It seems quite unlikely that humans happen to manifest a maximal level of general intelligence, even relative to the goals and environment for which they have been evolutionarily adapted.

There is also a common intuition in the AGI community that various real-world general intelligences will tend to share certain common properties; though there is less agreement on what these properties are!

1.2 The Core AGI Hypothesis

Another point broadly shared in the AGI community is confidence in what I would venture to call the “core AGI hypothesis,” i.e. that

Core AGI hypothesis: *the creation and study of synthetic intelligences with sufficiently broad (e.g. human-level) scope and strong generalization capability, is at bottom qualitatively different from the creation and study of synthetic intelligences with significantly narrower scope and weaker generalization capability.*

This “**core AGI hypothesis**” is explicitly articulated in English for the first time here in this review paper (it was presented previously in Japanese in Goertzel (2014)). I highlight it because it is something with which nearly all researchers in the AGI community agree, regardless of their different conceptualizations of the AGI concept and their different architectural, theoretical, technical and engineering approaches.⁹

If this core hypothesis is correct, then distinguishing AGI as a separate pursuit and system class and property from the “narrow AI” that has come to constitute the main stream of the AI field, is a sensible and productive thing to do.

Note, the core AGI hypothesis doesn’t imply there is zero commonality between narrower-scope AI work and AGI work. For instance, if a researcher is engineering a self-driving car via

8. Some researchers have suggested that cetacea might possess general intelligence comparable to that of humans, though very different in nature (Dye, 2010)

9. It must be admitted that this “core hypothesis”, as articulated here, is rather vague. More precise versions can be formulated, but then this seems to require making decisions that only a fraction of the AGI community will agree with. The reality is that currently the level of conceptual agreement among members of the AGI community pursuing different research approaches is mainly at the level of broad, vaguely-stated concepts, rather than precise formulations.

a combination of specialized AI techniques, they might use methods from the field of transfer learning (Taylor, Kuhlmann, and Stone, 2008) to help each component of the car’s control system (e.g. the object recognition system, the steering control system, etc.) better able to deal with various diverse situations it might encounter. This sort of transfer learning research, having to do with generalization, might have some overlap with the work one would need to do to make a generalized “AGI driver“ that could, on its own, adapt its operations flexibly from one vehicle or one environment to another. But the core AGI hypothesis proposes that, in order to make the latter sort of AGI driver, additional architectural and dynamical principles would be required, beyond those needed to aid in the human-mediated, machine learning aided creation of a variety of narrowly specialized AI driving systems.

1.3 The Scope of the AGI Field

Within the scope of the core AGI hypothesis, a number of different approaches to defining and characterizing AGI are under current study, encompassing psychological, mathematical, pragmatic and cognitive architecture perspectives. This paper surveys the contemporary AGI field in a fairly inclusive way. It also discusses the question of how much evidence exists for the core AGI hypothesis – and how the task of gathering more evidence about this hypothesis should best be pursued. The goal here is not to present any grand new conclusions, but rather to summarize and systematize some of the key aspects AGI as manifested in current science and engineering efforts.

It is argued here that most contemporary approaches to designing AGI systems fall into four top-level categories: symbolic, emergentist, hybrid and universalist. Leading examples of each category are provided, and the generally perceived pros and cons of each category are summarized.

Not all contemporary AGI approaches seek to create *human-like* general intelligence specifically. But it is argued here, that, for any approach which does, there is a certain set of key cognitive processes and interactions that it must come to grips with, including familiar constructs such as working and long-term memory, deliberative and reactive processing, perception, action and reinforcement learning, metacognition and so forth.

A robust theory of general intelligence, human-like or otherwise, remains elusive. Multiple approaches to defining general intelligence have been proposed, and in some cases these coincide with different approaches to designing AGI systems (so that various systems aim for general intelligence according to different definitions). The perspective presented here is that a mature theory of AGI would allow one to theoretically determine, based on a given environment and goal set and collection of resource constraints, the optimal AGI architecture for achieving the goals in the environments given the constraints. Lacking such a theory at present, researchers must conceive architectures via diverse theoretical paradigms and then evaluate them via practical metrics.

Finally, in order for a community to work together toward common goals, environments and metrics for evaluation of progress are necessary. Metrics for assessing the achievement of human-level AGI are argued to be fairly straightforward, including e.g. the classic Turing test, and the test of operating a robot that can graduate from elementary school or university. On the other hand, metrics for assessing partial progress toward, human-level AGI are shown to be more controversial and problematic, with different metrics suiting different AGI approaches, and with the possibility of systems whose partial versions perform poorly on commonsensical metrics, yet whose complete versions perform well. The problem of defining agreed-upon metrics for incremental progress

remains largely open, and this constitutes a substantial challenge for the young field of AGI moving forward.

2. Characterizing AGI and General Intelligence

One interesting feature of the AGI community, alluded to above, is that it does not currently agree on any single definition of the AGI concept – though there is broad agreement on the general intuitive nature of AGI, along the lines I’ve summarized above; and broad agreement that some form of the core AGI hypothesis is true. There is a mature theory of general intelligence in the psychology field, and a literature in the AGI field on the formal mathematical definition of intelligence; both of these will be reviewed below; however, none of the psychological nor mathematical conceptions of general intelligence are accepted as foundational in their details, by more than a small plurality of the AGI community. Rather, the formulation of a detailed and rigorous theory of “what AGI is”, is a small but significant part of the AGI community’s ongoing research. The bulk of the emerging AGI community’s efforts is devoted to devising and implementing designs for AGI systems, and developing theories regarding the best way to do so; but the fleshing out of the concept of “AGI” is being accomplished alongside and in synergy with these other tasks.

It must be noted, however, that the term “AI” also has many different meanings within the AI research community, with no clear agreement on the definition. George Luger’s popular AI textbook famously defined it as “that which AI practitioners do.” The border between AI and advanced algorithmics is often considered unclear. A common joke is that, as soon as a certain functionality has been effectively achieved by computers, it’s no longer considered AI. The situation with the ambiguity of “AGI” is certainly no worse than that with the ambiguity of the term “AI” itself.

In terms of basic semantics, the term “AGI” has been variously used to describe

- a property of certain systems (“AGI” as the intersection of “artificial” (i.e. synthetic) and “generally intelligent”)
- a system that displays this property (an “AGI” meaning “an AGI system”)
- the field of endeavor pursuing the creation of AGI systems, and the study of the nature of AGI

AGI is related to many other terms and concepts. Joscha Bach (Bach, 2009) has elegantly characterized it in terms of the quest to create “synthetic intelligence.” One also finds communities of researchers working toward AGI-related goals under the labels “computational intelligence”, “natural intelligence”, “cognitive architecture”, “biologically inspired cognitive architecture” (BICA), and many others. Each of these labels was introduced with a certain underlying purpose, and has a specific collection of concepts and approaches associated with it; each corresponds to a certain perspective or family of perspectives. The specific purpose underlying the concept and term “AGI” is to focus attention on the **general scope** and **generalization capability** of certain intelligent systems, such as humans, theoretical system like AIXI (Hutter, 2005), and a subset of potential future synthetic intelligences. That is, roughly speaking, an AGI system is a synthetic intelligence that has a general scope and is good at generalization across various goals and contexts.

The ambiguity of the concept of “AGI” relates closely to the underlying ambiguity of the concepts of “intelligence” and “general intelligence.” The AGI community has embraced, to varying

extents, a variety of characterizations of general intelligence, finding each of them to contribute different insights to the AGI quest. Legg and Hutter (Legg and Hutter, 2007a) wrote a paper summarizing and organizing over 70 different published definitions of “intelligence”, most oriented toward general intelligence, emanating from researchers in a variety of disciplines. In the rest of this section I will overview the main approaches to defining or characterizing general intelligence taken in the AGI field.

2.1 AGI versus Human-Level AI

One key distinction to be kept in mind as we review the various approaches to characterizing AGI, is the distinction between AGI and the related concept of “human-level AI” (which is usually used to mean, in effect: human-level, reasonably human-like AGI).

AGI is a fairly abstract notion, which is not intrinsically tied to any particular characteristics of human beings. Some properties of human general intelligence may in fact be universal among all powerful AGIs, but given our current limited understanding of general intelligence, it’s not yet terribly clear what these may be.

The concept of “human-level AGI”, interpreted literally, is confusing and ill-defined. It’s difficult to place the intelligences of all possible systems in a simple hierarchy, according to which the “intelligence level” of an arbitrary intelligence can be compared to the “intelligence level” of a human. Some researchers, as will be discussed below, have proposed universal intelligence measures that could be used in this way; but currently the details and utility of such measures are both quite contentious. To keep things simpler, here I will interpret “human-level AI” as meaning “human-level and roughly human-like AGI,” a restriction that makes the concept much easier to handle. For AGI systems that are supposed to operate in similar sorts of environments to humans, according to cognitive processes vaguely similar to those used by humans, the concept of “human level” is relatively easy to understand.

The concept of “AGI” appears more theoretically fundamental than “human-level AGI”; however, its very breadth can also be problematic. “Human-level AGI” is more concrete and specific, which lets one take it in certain directions more easily than can be done with general AGI. In our discussions on evaluations and metrics below, for example, we will restrict attention to human-level AGI systems, because otherwise creating metrics to compare qualitatively different AGI systems becomes a much trickier problem.

2.2 The Pragmatic Approach to Characterizing General Intelligence

The pragmatic approach to conceptualizing general intelligence is typified by the AI Magazine article “Human Level Artificial Intelligence? Be Serious!”, written by Nils Nilsson, one of the early leaders of the AI field (Nilsson, 2005). Nilsson’s view is

... that achieving real Human Level artificial intelligence would necessarily imply that most of the tasks that humans perform for pay could be automated. Rather than work toward this goal of automation by building special-purpose systems, I argue for the development of general-purpose, educable systems that can learn and be taught to perform any of the thousands of jobs that humans can perform. Joining others who have made similar proposals, I advocate beginning with a system that has minimal,

although extensive, built-in capabilities. These would have to include the ability to improve through learning along with many other abilities.

In this perspective, once an AI obsoletes humans in most of the practical things we do, it's got general Human Level intelligence. The implicit assumption here is that humans are the generally intelligent system we care about, so that the best practical way to characterize general intelligence is via comparison with human capabilities.

The classic Turing Test for machine intelligence – simulating human conversation well enough to fool human judges (Turing, 1950) – is pragmatic in a similar sense to Nilsson. But the Turing test has a different focus, on emulating humans. Nilsson isn't interested in whether an AI system can fool people into thinking it's a human, but rather in whether an AI system can do the useful and important practical things that people can do.

2.3 Psychological Characterizations of General Intelligence

The psychological approach to characterizing general intelligence also focuses on human-like general intelligence; but rather than looking directly at practical capabilities, it tries to isolate deeper underlying capabilities that enable these practical capabilities. In practice it encompasses a broad variety of sub-approaches, rather than presenting a unified perspective.

Viewed historically, efforts to conceptualize, define, and measure intelligence in humans reflect a distinct trend from general to specific (it is interesting to note the similarity between historical trends in psychology and AI) (Gregory, 2004). Thus, early work in defining and measuring intelligence was heavily influenced by Spearman, who in 1904 proposed the psychological factor *g* (the “*g* factor”, for general intelligence) (Spearman, 1904). Spearman argued that *g* was biologically determined, and represented the overall intellectual skill level of an individual. A related advance was made in 1905 by Binet and Simon, who developed a novel approach for measuring general intelligence in French schoolchildren (Binet and Simon, 1916). A unique feature of the Binet-Simon scale was that it provided comprehensive age norms, so that each child could be systematically compared with others across both age and intellectual skill level. In 1916, Terman introduced the notion of an intelligence quotient or IQ, which is computed by dividing the test-taker's mental age (i.e., their age-equivalent performance level) by their physical or chronological age (Terman, 1915).

In subsequent years, psychologists began to question the concept of intelligence as a single, undifferentiated capacity. There were two primary concerns. First, while performance within an individual across knowledge domains is somewhat correlated, it is not unusual for skill levels in one domain to be considerably higher or lower than in another (i.e., intra-individual variability). Second, two individuals with comparable overall performance levels might differ significantly across specific knowledge domains (i.e., inter-individual variability). These issues helped to motivate a number of alternative theories, definitions, and measurement approaches, which share the idea that intelligence is multifaceted and variable both within and across individuals. Of these approaches, a particularly well-known example is Gardner's theory of multiple intelligences, which proposes eight distinct forms or types of intelligence: (1) linguistic, (2) logical-mathematical, (3) musical, (4) bodily-kinesthetic, (5) spatial, (6) interpersonal, (7) intrapersonal, and (8) naturalist (Gardner, 1999). Gardner's theory suggests that each individual's intellectual skill is represented by an intelligence profile, that is, a unique mosaic or combination of skill levels across the eight forms of intelligence.

2.3.1 COMPETENCIES CHARACTERIZING HUMAN-LEVEL GENERAL INTELLIGENCE

Another approach to understanding general intelligence based on the psychology literature, is to look at the various *competencies* that cognitive scientists generally understand humans to display. The following list of competencies was assembled at the 2009 AGI Roadmap Workshop (Adams et al., 2012) via a group of 12 experts, including AGI researchers and psychologists, based on a review of the AI and psychology literatures. The list is presented as a list of broad areas of capability, each one then subdivided into specific sub-areas:

- Perception
 - Vision: image and scene analysis and understanding
 - Hearing: identifying the sounds associated with common objects; understanding which sounds come from which sources in a noisy environment
 - Touch: identifying common objects and carrying out common actions using touch alone
 - Crossmodal: Integrating information from various senses
 - Proprioception: Sensing and understanding what its body is doing
- Actuation
 - Physical skills: manipulating familiar and unfamiliar objects
 - Tool use, including the flexible use of ordinary objects as tools
 - Navigation, including in complex and dynamic environments
- Memory
 - Implicit: Memory the content of which cannot be introspected
 - Working: Short-term memory of the content of current/recent experience (awareness)
 - Episodic: Memory of a first-person experience (actual or imagined) attributed to a particular instance of the agent as the subject who had the experience
 - Semantic: Memory regarding facts or beliefs
 - Procedural: Memory of sequential/parallel combinations of (physical or mental) actions, often habituated (implicit)
 -
- Learning
 - Imitation: Spontaneously adopt new behaviors that the agent sees others carrying out
 - Reinforcement: Learn new behaviors from positive and/or negative reinforcement signals, delivered by teachers and/or the environment
 - Interactive verbal instruction
 - Learning from written media
 - Learning via experimentation

- Reasoning
 - Deduction, from uncertain premises observed in the world
 - Induction, from uncertain premises observed in the world
 - Abduction, from uncertain premises observed in the world
 - Causal reasoning, from uncertain premises observed in the world
 - Physical reasoning, based on observed “fuzzy rules” of naive physics
 - Associational reasoning, based on observed spatiotemporal associations
- Planning
 - Tactical
 - Strategic
 - Physical
 - Social
- Attention
 - Visual Attention within the agent’s observations of its environment
 - Social Attention
 - Behavioral Attention
- Motivation
 - Subgoal creation, based on the agent’s preprogrammed goals and its reasoning and planning
 - Affect-based motivation
 - Control of emotions
- Emotion
 - Expressing Emotion
 - Perceiving / Interpreting Emotion
- Modeling Self and Other
 - Self-Awareness
 - Theory of Mind
 - Self-Control
 - Other-Awareness
 - Empathy

- Social Interaction
 - Appropriate Social Behavior
 - Communication about and oriented toward social relationships
 - Inference about social relationships
 - Group interactions (e.g. play) in loosely-organized activities
- Communication
 - Gestural communication to achieve goals and express emotions
 - Verbal communication using natural language in its life-context
 - Pictorial Communication regarding objects and scenes with
 - Language acquisition
 - Cross-modal communication
- Quantitative
 - Counting sets of objects in its environment
 - Simple, grounded arithmetic with small numbers
 - Comparison of observed entities regarding quantitative properties
 - Measurement using simple, appropriate tools
- Building/Creation
 - Physical: creative constructive play with objects
 - Conceptual invention: concept formation
 - Verbal invention
 - Social construction (e.g. assembling new social groups, modifying existing ones)

Different researchers have different views about which of the above competency areas is most critical, and as you peruse the list, you may feel that it over or under emphasizes certain aspects of intelligence. But it seems clear that any software system that could flexibly and robustly display competency in all of the above areas, would be broadly considered a strong contender for possessing human-level general intelligence.

2.4 A Cognitive-Architecture Perspective on General Intelligence

Complementing the above perspectives, Laird et al (Laird et al., 2009) have composed a list of “requirements for human-level intelligence” from the standpoint of designers of cognitive architectures. Their own work has mostly involved the SOAR cognitive architecture, which has been pursued from the AGI perspective, but also from the perspective of accurately simulating human cognition:

- R0. FIXED STRUCTURE FOR ALL TASKS (i.e., explicit loading of knowledge files or software modification should not be done when the AGI system is presented with a new task)
- R1. REALIZE A SYMBOL SYSTEM (i.e., the system should be able to create symbolism and utilize symbolism internally, regardless of whether this symbolism is represented explicitly or implicitly within the system’s knowledge representation)
- R2. REPRESENT AND EFFECTIVELY USE MODALITY-SPECIFIC KNOWLEDGE
- R3. REPRESENT AND EFFECTIVELY USE LARGE BODIES OF DIVERSE KNOWLEDGE
- R4. REPRESENT AND EFFECTIVELY USE KNOWLEDGE WITH DIFFERENT LEVELS OF GENERALITY
- R5. REPRESENT AND EFFECTIVELY USE DIVERSE LEVELS OF KNOWLEDGE
- R6. REPRESENT AND EFFECTIVELY USE BELIEFS INDEPENDENT OF CURRENT PERCEPTION
- R7. REPRESENT AND EFFECTIVELY USE RICH, HIERARCHICAL CONTROL KNOWLEDGE
- R8. REPRESENT AND EFFECTIVELY USE META-COGNITIVE KNOWLEDGE
- R9. SUPPORT A SPECTRUM OF BOUNDED AND UNBOUNDED DELIBERATION (where “bounded” refers to computational space and time resource utilization)
- R10. SUPPORT DIVERSE, COMPREHENSIVE LEARNING
- R11. SUPPORT INCREMENTAL, ONLINE LEARNING

As Laird et al (Laird et al., 2009) note, there are no current AI systems that plainly fulfill all these requirements (although the precise definitions of these requirements may be open to a fairly broad spectrum of interpretations).

It is worth remembering, in this context, Stan Franklin’s careful articulation of the difference between a software “agent” and a mere “program” (Franklin and Graesser, 1997):

An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.

Laird and Wray’s requirements do not specify that the general intelligence must be an autonomous agent rather than a program. So, their requirements span both “agent AI” and “tool AI”. However, if we piece together Franklin’s definition with Laird and Wray’s requirements, we get a reasonable stab at a characterization of a “generally intelligent agent”, from the perspective of the cognitive architecture designer.

2.5 A Mathematical Approach to Characterizing General Intelligence

In contrast to approaches focused on human-like general intelligence, some researchers have sought to understand general intelligence *in general*. The underlying intuition here is that

- Truly, absolutely general intelligence would only be achievable given infinite computational ability. For any computable system, there will be some contexts and goals for which it's not very intelligent
- However, some finite computational systems will be more generally intelligent than others, and it's possible to quantify this extent

This approach is typified by the recent work of Legg and Hutter (Legg and Hutter, 2007b), who give a formal definition of general intelligence based on the Solomonoff-Levin prior. Put very roughly, they define intelligence as the average reward-achieving capability of a system, calculated by averaging over all possible reward-summable environments, where each environment is weighted in such a way that more compactly describable programs have larger weights.

According to this sort of measure, humans are nowhere near the maximally generally intelligent system. However, humans are more generally intelligent than, say, rocks or worms.¹⁰

While the original form of Legg and Hutter's definition of intelligence is impractical to compute, a more tractable approximation has recently been developed (Legg and Veness, 2013). Also, Achler (Achler, 2012b) has proposed an interesting, pragmatic AGI intelligence measurement approach explicitly inspired by these formal approaches, in the sense that it explicitly balances the effectiveness of a system at solving problems with the compactness of its solutions. This is similar to a common strategy in evolutionary program learning, where one uses a fitness function comprising an accuracy term and an "Occam's Razor" compactness term.

2.6 The Adaptationist Approach to Characterizing General Intelligence

Another perspective views general intelligence as closely tied to the environment in which it exists. Pei Wang has argued carefully for a conception of general intelligence as adaptation to the environment using insufficient resources (Wang, 2006). A system may be said to have greater general intelligence, if it can adapt effectively to a more general class of environments, within reapeilistic resource constraints.

In a 2010 paper, I sought to modify Legg and Hutter's mathematical approach in an attempt to account for the factors Wang's definition highlights (Goertzel, 2010);

- The *pragmatic general intelligence* is defined relative to a given probability distribution over environments and goals, as the average goal-achieving capability of a system, calculated by

10. A possible practical issue with this approach is that the quantitative general-intelligence values it yields are dependent on the choice of reference Universal Turing Machine underlying the measurement of program length. A system is judged as intelligent largely based on extent that it solves simple problems effectively, but the definition of "simple", in practice, depends on the assumed UTM – what is simple to one UTM may be complex to another. In the limit of infinitely large problems, this issue goes away due to the ability of any UTM to simulate any other one, but human intelligence is not currently mainly concerned with the limit of infinitely large problems. This means that in order to turn these ideas into a practical intelligence measure, one would have to make a commitment to a particular UTM; and current science and philosophy don't give strong guidance regarding which one to choose. How large a difficulty this constitutes in practice remains to be seen. Researchers working on this sort of approach, tend not to consider this a real problem.

weighted-averaging over all possible environments and goals, using the given distribution to determine the weights

- The *generality* of a system’s intelligence is defined in a related way, as (roughly speaking) the entropy of the class of environments over which the system displays high pragmatic general intelligence
- The *efficient pragmatic general intelligence* is defined relative to a given probability distribution over environments and goals, as the average effort-normalized goal-achieving capability of a system, calculated by weighted-averaging over all possible environments and goals, using the given distribution to determine the weights. The effort-normalized goal-achieving capability of a system is defined by taking its goal-achieving capability (relative to a particular goal and environment), and dividing it by the computational effort the system must expend to achieve that capability

Lurking in this vicinity are some genuine differences of perspective with the AGI community, regarding the proper way to conceive general intelligence. Some theorists (e.g. Legg and Hutter) argue that intelligence is purely a matter of capability, and that the intelligence of a system is purely a matter of its behaviors, and is independent of how much effort it expends in achieving its behaviors. On the other hand, some theorists (e.g. Wang) believe that the essence of general intelligence lies in the complex systems of compromises needed to achieve a reasonable degree of generality of adaptation using limited computational resources. In the latter, adaptationist view, the sorts of approaches to goal-achievement that are possible in the theoretical case of infinite or massive computational resources, have little to do with real-world general intelligence. But in the former view, real-world general intelligence can usefully be viewed as a modification of infinite-resources, infinitely-general intelligence to the case of finite resources.

2.7 The Embodiment Focused Approach to Characterizing General Intelligence

A close relative of the adaptationist approach, but with a very different focus that leads to some significant conceptual differences as well, is what we may call the **embodiment approach** to characterizing general intelligence. In brief this perspective holds that intelligence is something that physical bodies do in physical environments. It holds that intelligence is best understood via focusing on the modulation of the body-environment interaction that an embodied system carries out as it goes about in the world. Rodney Brooks is one of the better known advocates of this perspective (Brooks, 2002).

Pfeifer and Bongard summarize the view of intelligence underlying this perspective adroitly as follows: “In spite of all the difficulties of coming up with a concise definition, and regardless of the enormous complexities involved in the concept of intelligence, it seems that whatever we intuitively view as intelligent is always vested with two particular characteristics: compliance and diversity. In short, intelligent agents always comply with the physical and social rules of their environment, and exploit those rules to produce diverse behavior.” (Pfeifer and Bongard, 2007). For example, they note: “All animals, humans and robots have to comply with the fact that there is gravity and friction, and that locomotion requires energy... [A]dapting to these constraints and exploiting them in particular ways opens up the possibility of walking, running, drinking from a cup, putting dishes on a table, playing soccer, or riding a bicycle.”

Pfeifer and Bonard go so far as to assert that intelligence, in the perspective they analyze it, doesn't apply to conventional AI software programs. "We ascribe intelligence only to ... real physical systems whose behavior can be observed as they interact with the environment. Software agents, and computer programs in general, are disembodied, and many of the conclusions drawn ... do not apply to them." Of course, this sort of view is quite contentious, and e.g. Pei Wang has argued against it in a paper titled "Does a Laptop Have a Body?" (Wang, 2009) – the point being that any software program with any kind of user interface is interacting with the physical world via some kind of body, so the distinctions involved are not as sharp as embodiment-oriented researchers sometimes imply.

Philosophical points intersect here with issues regarding research focus. Conceptually, the embodiment perspective asks whether it even make sense to talk about human-level or human-like AGI in a system that lacks a vaguely human-like body. Focus-wise, this perspective suggests that, if one is interested in AGI, it makes sense to put resources on achieving human-like intelligence the way evolution did, i.e. in the context of controlling a body with complex sensors and actuators in a complex physical world.

The overlap between the embodiment and adaptationist approaches is strong, because historically, human intelligence evolved specifically to adapt to the task of controlling a human body in certain sorts of complex environment, given limited energetic resources and subject to particular physical constraints. But, the two approaches are not identical, because the embodiment approach posits that adaptation to physical body-control tasks under physical constraints is key, whereas the adaptationist approach holds that the essential point is more broadly-conceived adaptation to environments subject to resource constraints.

3. Approaches to Artificial General Intelligence

As appropriate for an early-stage research field, there is a wide variety of different approaches to AGI in play. Fairly comprehensive reviews have been provided by Wlodek Duch's review paper from the AGI-08 conference (Duch, Oentaryo, and Pasquier, 2008); and Alexei Samsonovich's BICA review paper (Samsonovich, 2010), which compares a number of (sometimes quite loosely) biologically inspired cognitive architectures in terms of a feature checklist, and was created collaboratively with the creators of the architectures. Hugo de Garis and I also wrote two review papers, one focused on biologically-inspired cognitive architectures (Goertzel et al., 2010a) and the other on computational neuroscience systems with AGI ambitions (De Garis et al., 2010). Here I will not try to review the whole field in detail; I will be content with describing the main categories of approaches, and briefly citing a few illustrative examples of each one.¹¹

Duch's survey (Duch, Oentaryo, and Pasquier, 2008), divides existing approaches into three paradigms – symbolic, emergentist and hybrid. Whether this trichotomy has any fundamental significance is somewhat contentious, but it is convenient given the scope of approaches currently and historically pursued, so I will use it to help structure the present brief review of AGI approaches. But I will deviate from Duch in a couple ways: I add one additional category ("universalist"), and I split the emergentist category into multiple subcategories.

11. The choice to include approach X here and omit approach Y, should not be construed as implying that I think X has more potential than Y, or even that X illustrates the category in which I've placed it better than Y would. Rather, there are a lot of AGI approaches and systems out there, and I've selected a few reasonably representative ones to give an overall picture of the field.

3.1 Symbolic AGI Approaches

A venerable tradition in AI focuses on the physical symbol system hypothesis (Nilsson, 2007), which states that minds exist mainly to manipulate symbols that represent aspects of the world or themselves. A physical symbol system has the ability to input, output, store and alter symbolic entities, and to execute appropriate actions in order to reach its goals. Generally, symbolic cognitive architectures focus on “working memory” that draws on long-term memory as needed, and utilize a centralized control over perception, cognition and action. Although in principle such architectures could be arbitrarily capable (since symbolic systems have universal representational and computational power, in theory), in practice symbolic architectures tend to be weak in learning, creativity, procedure learning, and episodic and associative memory. Decades of work in this tradition have not compellingly resolved these issues, which has led many researchers to explore other options.

Perhaps the most impressive successes of symbolic methods on learning problems have occurred in the areas of Genetic Programming (GP) (Koza, 1992), Inductive Logic Programming (Muggleton, 1991), and probabilistic learning methods such as Markov Logic Networks (MLN) (Richardson and Domingos, 2006). These techniques are interesting from a variety of theoretical and practical standpoints. For instance, it is notable that, GP and MLN have been usefully applied to high-level symbolic relationships, and also to quantitative data resulting directly from empirical observations, depending on how one configures them and how one prepares their inputs. Another important observation one may make about these methods is that, in each case, the ability to do data-driven learning using an underlying symbolic representation, comes along with a lack of transparency in how and why the learning algorithms come up with the symbolic constructs that they do. Nontrivially large GP program trees are generally quite opaque to the human reader, though in principle using a comprehensible symbolic formalism. The propositions making up a Markov Logic Network are easy to understand, but the reasons that MLN weight learning ranks one propositional rule higher than another over a given set of evidence, are obscure and not easily determinable from the results MLN produces. In some ways these algorithms blur the border between symbolic and subsymbolic, because they use underlying symbolic representation languages according to algorithms that produce large, often humanly inscrutable combinations of data elements in a manner conceptually similar to many subsymbolic learning algorithms.

Indeed, the complex, somewhat “emergentist” nature of “symbolic” algorithms like GP and MLN provides a worthwhile reminder that the “symbolic vs. subsymbolic” dichotomy, while heuristically valuable for describing the AI and AGI approaches existent at the current time, is not necessarily a clear, crisp, fundamentally grounded distinction. It is utilized here more for its sociological descriptive value, as for its core value as a scientific, mathematical or philosophical distinction.

A few illustrative symbolic cognitive architectures are:

- **ACT-R** (Anderson and Lebiere, 2003) is fundamentally a symbolic system, but Duch classifies it as a hybrid system because it incorporates connectionist-style activation spreading in a significant role; and there is an experimental thoroughly connectionist implementation to complement the primary mainly-symbolic implementation. Its combination of SOAR-style “production rules” with large-scale connectionist dynamics allows it to simulate a variety of human psychological phenomena.

- **Cyc** (Lenat and Guha, 1989) is an AGI architecture based on predicate logic as a knowledge representation, and using logical reasoning techniques to answer questions and derive new knowledge from old. It has been connected to a natural language engine, and designs have been created for the connection of Cyc with Albus's 4D-RCS (Albus, 2001). Cyc's most unique aspect is the large database of commonsense knowledge that Cycorp has accumulated (millions of pieces of knowledge, entered by specially trained humans in predicate logic format); part of the philosophy underlying Cyc is that once a sufficient quantity of knowledge is accumulated in the knowledge base, the problem of creating human-level general intelligence will become much less difficult due to the ability to leverage this knowledge.
- **EPIC** (Rosbe, Chong, and Kieras, 2001), a cognitive architecture aimed at capturing human perceptual, cognitive and motor activities through several interconnected processors working in parallel. The system is controlled by production rules for cognitive processor and a set of perceptual (visual, auditory, tactile) and motor processors operating on symbolically coded features rather than raw sensory data. It has been connected to SOAR for problem solving, planning and learning.
- **ICARUS** (Langley, 2005), an integrated cognitive architecture for physical agents, with knowledge specified in the form of reactive skills, each denoting goal-relevant reactions to a class of problems. The architecture includes a number of modules: a perceptual system, a planning system, an execution system, and several memory systems.
- **SNePS** (Semantic Network Processing System) (Shapiro et al., 2007) is a logic, frame and network-based knowledge representation, reasoning, and acting system that has undergone over three decades of development, and has been used for some interesting prototype experiments in language processing and virtual agent control.
- **SOAR** (Laird, 2012), a classic example of expert rule-based cognitive architecture. designed to model general intelligence. It has recently been extended to handle sensorimotor functions and reinforcement learning.

A caricature of some common attitudes for and against the symbolic approach to AGI would be:

- **For:** Symbolic thought is what most strongly distinguishes humans from other animals; it's the crux of human general intelligence. Symbolic thought is precisely what lets us generalize most broadly. It's possible to realize the symbolic core of human general intelligence independently of the specific neural processes that realize this core in the brain, and independently of the sensory and motor systems that serve as (very sophisticated) input and output conduits for human symbol-processing.
- **Against:** While these symbolic AI architectures contain many valuable ideas and have yielded some interesting results, they seem to be incapable of giving rise to the emergent structures and dynamics required to yield humanlike general intelligence using feasible computational resources. Symbol manipulation emerged evolutionarily from simpler processes of perception and motivated action; and symbol manipulation in the human brain emerges from these same sorts of processes. Divorcing symbol manipulation from the underlying substrate

of perception and motivated action doesn't make sense, and will never yield generally intelligent agents, at best only useful problem-solving tools.

3.2 Emergentist AGI Approaches

Another species of AGI design expects abstract symbolic processing – along with every other aspect of intelligence – to emerge from lower-level “subsymbolic” dynamics, which sometimes (but not always) are designed to simulate neural networks or other aspects of human brain function. Today's emergentist architectures are sometimes very strong at recognizing patterns in high-dimensional data, reinforcement learning and associative memory; but no one has yet compellingly shown how to achieve high-level functions such as abstract reasoning or complex language processing using a purely subsymbolic, emergentist approach. There are research results doing inference and language processing using subsymbolic architectures, some of which are reviewed in (Hammer and Hitzler, 2007); but these mainly involve relatively simplistic problem cases. The most broadly effective reasoning and language processing systems available are those utilizing various forms of symbolic representations, though often also involving forms of probabilistic, data-driven learning, as in examples like Markov Logic Networks (Richardson and Domingos, 2006) and statistical language processing (Jurafsky and James, 2000).

A few illustrative subsymbolic, emergentist cognitive architectures are:

- **DeSTIN** (Arel, Rose, and Karnowski, 2009; Arel, Rose, and Coop, 2009) is a hierarchical temporal pattern recognition architecture, with some similarities to HTM (Hawkins and Blakeslee, 2007) but featuring more complex learning mechanisms. It has been integrated into the CogPrime (Goertzel et al., 2011) architecture to serve as a perceptual subsystem; but is primarily being developed to serve as the center of its own AGI design, assisted via action and reinforcement hierarchies.
- **Hierarchical Temporal Memory (HTM)** (Hawkins and Blakeslee, 2007) is a hierarchical temporal pattern recognition architecture, presented as both an AI / AGI approach and a model of the cortex. So far it has been used exclusively for vision processing, but a conceptual framework has been outlined for extension to action and perception/action coordination.
- **SAL** (Jilk and Lebiere, 2008), based on the earlier and related **IBCA** (Integrated Biologically-based Cognitive Architecture) is a large-scale emergent architecture that seeks to model distributed information processing in the brain, especially the posterior and frontal cortex and the hippocampus. So far the architectures in this lineage have been used to simulate various human psychological and psycholinguistic behaviors, but haven't been shown to give rise to higher-level behaviors like reasoning or subgoaling.
- **NOMAD** (Neurally Organized Mobile Adaptive Device) automata and its successors (Krichmar and Edelman, 2006) are based on Edelman's “Neural Darwinism” model of the brain, and feature large numbers of simulated neurons evolving by natural selection into configurations that carry out sensorimotor and categorization tasks. This work builds conceptually on prior work by Edelman and colleagues on the “Darwin” series of brain-inspired perception systems (Reeke Jr, Sporns, and Edelman, 1990).
- Ben Kuipers and his colleagues (Modayil and Kuipers, 2007; Mugan and Kuipers, 2008, 2009) have pursued an extremely innovative research program which combines qualitative

reasoning and reinforcement learning to enable an intelligent agent to learn how to act, perceive and model the world. Kuipers’ notion of “bootstrap learning” involves allowing the robot to learn almost *everything* about its world, including for instance the structure of 3D space and other things that humans and other animals obtain via their genetic endowments.

- Tsvi Achler (Achler, 2012b) has demonstrated neural networks whose weights adapt according to a different methodology than the usual, combining feedback and feedforward dynamics in a particular way, with the result that the weights in the network have a clear symbolic meaning. This provides a novel approach to bridging the symbolic-subsymbolic gap.

There has also been a great deal of work relevant to these sorts of architectures, done without explicit reference to cognitive architectures, under labels such as “deep learning” – e.g. Andrew Ng’s well known work applying deep learning to practical vision processing problems (Socher et al., 2012; Le, 2013), and the work of Tommaso Poggio and his team which achieves deep learning via simulations of visual cortex (Anselmi et al., 2013). And there is a set of emergentist architectures focused specifically on developmental robotics, which we will review below in a separate subsection, as all of these share certain common characteristics.

A caricature of some common attitudes for and against the emergentist approach to AGI would be:

- **For:** The brain consists of a large set of simple elements, complexly self-organizing into dynamical structures in response to the body’s experience. So, the natural way to approach AGI is to follow a similar approach: a large set of simple elements capable of appropriately adaptive self-organization. When a cognitive faculty is achieved via emergence from subsymbolic dynamics, then it automatically has some flexibility and adaptiveness to it (quite different from the “brittleness” seen in many symbolic AI systems). The human brain is actually very similar to the brains of other mammals, which are mostly involved in processing high-dimensional sensory data and coordinating complex actions; this sort of processing, which constitutes the foundation of general intelligence, is most naturally achieved via subsymbolic means.
- **Against:** The brain happens to achieve its general intelligence via self-organizing networks of neurons, but to focus on this underlying level is misdirected. What matters is the cognitive “software” of the mind, not the lower-level hardware or wetware that’s used to realize it. The brain has a complex architecture that evolution has honed specifically to support advanced symbolic reasoning and other aspects of human general intelligence; what matters for creating human-level (or greater) intelligence is having the right information processing architecture, not the underlying mechanics via which the architecture is implemented.

3.2.1 COMPUTATIONAL NEUROSCIENCE AS A ROUTE TO AGI

One commonsensical approach to AGI, falling conceptually under the “emergentist” umbrella, would be to use *computational neuroscience* to create a model of how the brain works, and then to use this model as an AGI system. If we understood the brain more fully, this would be an extremely effective approach to creating the world’s first human-level AGI. Given the reality of our currently limited understanding of the brain and how best to digitally simulate it, the computational

neuroscience approach to AGI is no panacea, and in fact is almost impossible to pursue – but this is an interesting direction nonetheless.

To understand the difficulty of taking this approach to AGI, consider some illustrative examples of contemporary large-scale computational neuroscience projects:

- Markram’s IBM “Blue Brain Project” , which used a “Blue Gene” supercomputer to simulate (at ion channel level of detail) the neural signaling of a cortical column of the rat brain. The long-term goal of the project, now continuing in the EU with a large sum of government funding under the label “Human Brain Project”, is to “be able to simulate the full cortex of the human brain” (Markram, 2006).
- Modha’s IBM “Cognitive Computation Project”, aimed at “reverse engineering the structure, function, dynamics and behavior of the human brain, and then delivering it in a small compact form factor consuming very low power that rivals the power consumption of the human brain.” The best publicized achievement of Modha’s team has been a simulation (at a certain level of accuracy) of a neural network the size of the “cortex of a cat”, with 10^9 neurons, and 10^{13} synapses (Frye, Ananthanarayanan, and Modha, 2007).
- Boahen’s “Neurogrid Project” (at Stanford), involving the creation of custom integrated circuits that emulate the way neurons compute. So far his “neuromorphic engineering” research group has built a silicon retina, intended to be developed into something capable of giving the blind some degree of sight; and a self organizing chip, that emulates the way a developing brain wires itself up (Silver et al., 2007).
- Horwitz’s “Large-Scale Brain Modeling” (at the US NIH) initiative, involving simulation of the dynamic assemblage of neural subnetworks performing cognitive tasks, especially those associated with audition and language, and with an emphasis on the alteration of these networks during brain disorders. Horwitz’s simulation work is guided closely by data gathered from brain imaging using fMRI, PET, and MEG (Horwitz, Friston, and Taylor, 2000).
- Izhikevich’s and Edelman’s “Large Scale Model of Thalamocortical Systems” , a simulation on a scale similar to that of the full human brain itself. By simulating the spiking and plasticity features of the neural cortex, they managed to reproduce certain special features of the brain, such as initial states sensitivity, brain wave propagation, etc. Their model was used to simulate a million spiking neurons consisting of multiple compartments, joined by a half billion synapses, with responses calibrated to reproduce known types of responses recorded in vitro in rats. In this simulation, they observed a variety of interesting phenomena, including: spontaneous activity, the emergence of waves and rhythms, and functional connectivity on different scales.(Izhikevich and Edelman, 2008). Izhikevich’s current proprietary work in his firm “The Brain Corporation” is founded on similar principles.
- Just’s “4CAPS” (Cortical Capacity-Constrained Concurrent Activation-based Production System) cognitive architecture, a hybrid of a computational neuroscience model and a symbolic AI system, intended to explain both behavioral and neuroimaging data. The architecture includes computational features such as variable-binding and constituent-structured representations, alongside more standard neural net structures and dynamics (Just and Varma, 2007).

These are all fantastic projects; however, they embody a broad scope of interpretations of the notion of “simulation” itself. Different researchers are approaching the task of large-scale brain simulation with very different objectives in mind, e.g.

1. Creating models that can actually be connected to parts of the human brain or body, and can serve the same role as the brain systems they simulate. (e.g. Boahen’s artificial cochlea and retina (Silver et al., 2007)).
2. Creating a precise functional simulation of a brain subsystem, i.e. one that simulates the subsystem’s internal dynamics and its mapping of inputs to outputs with adequate fidelity to explain exactly what the brain subsystem does to control the organism. (something that so far has been done compelling only on a small scale for very specialized brain systems; Horwitz’s work is pushing in this direction on a somewhat larger scale than typical).
3. Creating models that quantitatively simulate the generic behavior and internal dynamics of a certain subsystem of the brain, but without precisely functionally simulating that subsystem. (e.g. Izhikevich and Edelman’s large-scale simulation, and Markram’s “statistically accurate” simulated cortical column).
4. Creating models that qualitatively simulate brain subsystems or whole brains at a high level, without simulating the particular details of dynamics or I/O, but with a goal of exploring some of the overall properties of the system. (e.g. Just’s 4CAPS work).
5. Creating models that demonstrate the capacity of hardware to simulate large neural models based on particular classes of equations, but without any claims about the match of the models in question to empirical neuroscience data. (e.g. Modha’s “cat” simulation).

All of the above are validly called “large scale brain simulations”, yet they constitute very different forms of research. Simulations in the first and fifth category are adequate to serve as components of AGI systems. Simulations in the other categories are useful for guiding neuroscience or hardware development, but are less directly useful for AGI.

Now, any one of these simulations, if advanced a little further in the right direction, could become more robustly functional and hence more clearly “AGI” rather than just computational neuroscience. But at the present time, our understanding of neuroscience isn’t quite advanced enough to guide the creation of computational neuroscience systems that actually display interesting intelligent behaviors, while still displaying high neural fidelity in their internal structures and dynamics. The bottleneck here isn’t really the computational simulation side, but more the neuroscience side – we just haven’t gathered the neuroscience data needed to spawn the creation of the neuroscience knowledge and understanding we’d need to drive this sort of AGI approach effectively yet.

Summing up, a caricature of some common attitudes for and against computational neuroscience as an approach to AGI would be:

- **For:** The brain is the only example we have of a system with a high level of general intelligence. So, emulating the brain is obviously the most straightforward path to achieving AGI. Neuroscience is advancing rapidly, and so is computer hardware; so, putting the two together, there’s a fairly direct path toward AGI by implementing cutting-edge neuroscience

models on massively powerful hardware. Once we understand how brain-based AGIs work, we will likely then gain the knowledge to build even better systems.

- **Against:** Neuroscience is advancing rapidly but is still at a primitive stage; our knowledge about the brain is extremely incomplete, and we lack understanding of basic issues like how the brain learns or represents abstract knowledge. The brain’s cognitive mechanisms are well-tuned to run efficiently on neural wetware, but current computer hardware has very different properties; given a certain fixed amount of digital computing hardware, one can create vastly more intelligent systems via crafting AGI algorithms appropriate to the hardware than via trying to force algorithms optimized for neural wetware onto a very different substrate.

3.2.2 ARTIFICIAL LIFE AS A ROUTE TO AGI

Another potential emergentist approach to AGI is to simulate a different type of biology: not the brain, but the evolving ecosystem that gave rise to the brain in the first place. That is: to seek AGI via artificial life ¹². Although Alife itself is a flourishing field, the artificial organisms created so far have been quite simplistic, more like simplified bugs or microscopic organisms than like creatures typically thought of as displaying a high level of general intelligence. Further, given the state of the art, each Alife simulation tends to reach an upper limit of complexity relatively soon; no one has yet managed to emulate the open-ended nature of biological ecosystems. Bruce Damer’s Evogrid (Damer et al., 2010) attempts to break through this logjam directly, via a massive distributed-computing powered use of chemistry simulations, in which evolutionary algorithms are used in an effort to evolve the best possible chemical soups; but this is still early-stage, though initial results are promising.

The main limitation of this approach is computational resource related: An ecosystem obviously requires a lot more computing resources than an individual brain or body. At present it’s unclear whether we have sufficient computational resources to realize individual human-level minds at feasible cost; simulating a whole ecosystem may be out of reach until a few more Moore’s Law doublings have occurred. Although, this isn’t a definitive objection, because it may be possible to craft artificial life-forms making exquisitely efficient use of digital computer architecture, or even of quantum computers or other radical new computing fabrics. At any rate, the Alife approach is not a major force in the AGI community at present, but it may surge as readily available computational power increases.

3.2.3 DEVELOPMENTAL ROBOTICS

Finally, one subset of emergentist cognitive architectures that I consider particularly important is the *developmental robotics* architectures, focused on controlling robots without significant “hard-wiring” of knowledge or capabilities, allowing robots to learn (and learn how to learn etc.) via their engagement with the world. A significant focus is often placed here on “intrinsic motivation,” wherein the robot explores the world guided by internal goals like novelty or curiosity, forming a model of the world as it goes along, based on the modeling requirements implied by its goals. Many of the foundations of this research area were laid by Juergen Schmidhuber’s work in the 1990s (Schmidhuber, 1991b,a, 1995, 2003), but now with more powerful computers and robots the area is leading to more impressive practical demonstrations.

12. See the site of the Society for Artificial Life <http://alife.org>

I mention here a handful of the illustrative initiatives in this area:

- Juyang Weng’s Dav (Han et al., 2002) and SAIL (Weng et al., 2000) projects involve mobile robots that explore their environments autonomously, and learn to carry out simple tasks by building up their own world-representations through both unsupervised and teacher-driven processing of high-dimensional sensorimotor data. The underlying philosophy is based on human child development (Weng and Hwang, 2006), the knowledge representations involved are neural network based, and a number of novel learning algorithms are involved, especially in the area of vision processing.
- FLOWERS (Baranès and Oudeyer, 2009), an initiative at the French research institute INRIA, led by Pierre-Yves Oudeyer, is also based on a principle of trying to reconstruct the processes of development of the human child’s mind, spontaneously driven by intrinsic motivations. Kaplan (Kaplan, 2008) has taken this project in a practical direction via the creation of a “robot playroom.” Experiential language learning has also been a focus of the project (Oudeyer and Kaplan, 2006), driven by innovations in speech understanding.
- IM-CLEVER¹³, a new European project coordinated by Gianluca Baldassarre and conducted by a large team of researchers at different institutions, which is focused on creating software enabling an iCub (Metta et al., 2008) humanoid robot to explore the environment and learn to carry out human childlike behaviors based on its own intrinsic motivations.

A caricature of some common attitudes for and against the developmental robotics approach to AGI would be:

- **For:** Young human children learn, mostly, by unsupervised exploration of their environment – using body and mind together to adapt to the world, with progressively increasing sophistication. This is the only way that we know of, for a mind to move from ignorance and incapability to knowledge and capability.
- **Against:** Robots, at this stage in the development of technology, are extremely crude compared to the human body, and thus don’t provide an adequate infrastructure for mind/body learning of the sort a young human child does. Due to the early stage of robotics technology, robotics projects inevitably become preoccupied with robotics particulars, and never seem to get to the stage of addressing complex cognitive issues. Furthermore, it’s unclear whether detailed sensorimotor grounding is actually necessary in order to create an AGI doing human-level reasoning and learning.

3.3 Hybrid AGI Architectures

In response to the complementary strengths and weaknesses of the symbolic and emergentist approaches, in recent years a number of researchers have turned to integrative, hybrid architectures, which combine subsystems operating according to the two different paradigms. The combination may be done in many different ways, e.g. connection of a large symbolic subsystem with a large subsymbolic system, or the creation of a population of small agents each of which is both symbolic and subsymbolic in nature.

13. <http://im-clever.noze.it/project/project-description>

Nils Nilsson expressed the motivation for hybrid AGI systems very clearly in his article at the AI-50 conference (which celebrated the 50th anniversary of the AI field) (Nilsson, 2007). While affirming the value of the “Physical Symbol System Hypothesis” (PSSH) that underlies classical symbolic AI, he argues that

the PSSH explicitly assumes that, whenever necessary, symbols will be grounded in objects in the environment through the perceptual and effector capabilities of a physical symbol system.

Thus, he continues,

I grant the need for non-symbolic processes in some intelligent systems, but I think they supplement rather than replace symbol systems. I know of no examples of reasoning, understanding language, or generating complex plans that are best understood as being performed by systems using exclusively non-symbolic processes....

AI systems that achieve human-level intelligence will involve a combination of symbolic and non-symbolic processing.

Hybrid architectures are often designed to leverage (hypothesized or empirically observed) “whole is greater than the sum of the parts” phenomena arising when multiple components are appropriately connected. This is philosophically related to the emergence phenomena at the conceptual heart of many subsymbolic architectures. In (Goertzel et al., 2011) the concept of “cognitive synergy” is formulated to capture this idea; it is conjectured that human-level AGI intrinsically depends on the synergetic interaction of multiple components (for instance, as in the CogPrime design (Goertzel et al., 2011), multiple memory systems each supplied with its own learning process).

A few illustrative hybrid cognitive architectures are:

- **CLARION** (Sun and Zhang, 2004) is a hybrid architecture that combines a symbolic component for reasoning on “explicit knowledge” with a connectionist component for managing “implicit knowledge.” Learning of implicit knowledge may be done via neural net, reinforcement learning, or other methods. The integration of symbolic and subsymbolic methods is powerful, but a great deal is still missing such as episodic knowledge and learning and creativity. Learning in the symbolic and subsymbolic portions is carried out separately rather than dynamically coupled.
- **CogPrime** (Goertzel et al., 2011), an AGI approach developed by myself and my colleagues, and being implemented within the OpenCog open source AI software platform. CogPrime integrates multiple learning algorithms associated with different memory types, using a weighted labeled hypergraph knowledge representation and making heavy use of probabilistic semantics. The various algorithms are designed to display “cognitive synergy” and work together to achieve system goals. It is currently being used to control video game characters, and a project to use it to control humanoid robots is in the planning stage.
- **DUAL** (Nestor and Kokinov, 2004) is arguably the most impressive system to come out of Marvin Minsky’s “Society of Mind” paradigm. It features a population of agents, each of which combines symbolic and connectionist representation, utilizing population-wide

self-organization to collectively carry out tasks such as perception, analogy and associative memory.

- **LIDA** (Franklin et al., 2012) is a comprehensive cognitive architecture heavily based on Bernard Baars' "Global Workspace Theory" (Baars and Franklin, 2009). It articulates a "cognitive cycle" integrating various forms of memory and intelligent processing in a single processing loop. The architecture ties in well with both neuroscience and cognitive psychology, but it deals most thoroughly with "lower level" aspects of intelligence; the handling of more advanced aspects like language and reasoning in LIDA has not yet been worked out in detail.
- **MicroPsi** (Bach, 2009) is an integrative architecture based on Dietrich Dorner's Psi model of motivation, emotion and intelligence. It has been tested on some practical control applications, and also on simulating artificial agents in a simple virtual world. MicroPsi's basis in neuroscience and psychology are extensive and carefully-drawn. Similar to LIDA, MicroPsi currently focuses on the "lower level" aspects of intelligence, not yet directly handling advanced processes like language and abstract reasoning.
- **PolyScheme** (Cassimatis, 2007) integrates multiple methods of representation, reasoning and inference schemes for general problem solving. Each Polyscheme "specialist" models a different aspect of the world using specific representation and inference techniques, interacting with other specialists and learning from them. Polyscheme has been used to model infant reasoning including object identity, events, causality, spatial relations.
- **Shruti** (Shastri and Ajjanagadde, 1993) is a biologically-inspired model of human reflexive inference, which uses a connectionist architecture to represent relations, types, entities and causal rules using focal-clusters.
- James Albus's **4D/RCS** robotics architecture shares a great deal with some of the emergentist architectures discussed above, e.g. it has the same hierarchical pattern recognition structure as DeSTIN and HTM, and the same three cross-connected hierarchies as DeSTIN, and shares with the developmental robotics architectures a focus on real-time adaptation to the structure of the world. However, 4D/RCS is not foundationally learning-based but relies on hard-wired architecture and algorithms, intended to mimic the qualitative structure of relevant parts of the brain (and intended to be *augmented* by learning, which differentiates it from emergentist approaches).

The nature of integration between components varies among the hybrid architectures. Some of them are in essence, multiple, disparate algorithms carrying out separate functions, encapsulated in black boxes and communicating results with each other. For instance, PolyScheme, ACT-R and CLARION all display this "modularity" property to a significant extent. On the other hand, architectures such as CogPrime, DUAL, Shruti, LIDA and MicroPsi feature richer integration – which makes their dynamics more challenging to understand and tune.

A caricature of some common attitudes for and against the hybrid approach to AGI would be:

- **For:** The brain is a complex system with multiple different parts, architected according to different principles but all working closely together; so in that sense, the brain is a

hybrid system. Different aspects of intelligence work best with different representational and learning mechanisms. If one designs the different parts of a hybrid system properly, one can get the different parts to work together synergetically, each contributing its strengths to help overcome the others' weaknesses. Biological systems tend to be messy, complex and integrative; searching for a single "algorithm of general intelligence" is an inappropriate attempt to project the aesthetics of physics or theoretical computer science into a qualitative different domain.

- **Against:** Gluing together a bunch of inadequate systems isn't going to make an adequate system. The brain uses a unified infrastructure (a neural network) for good reason; when you try to tie together qualitatively different components, you get a brittle system that can't adapt that well, because the different components can't work together with full flexibility. Hybrid systems are inelegant, and violate the "Occam's Razor" heuristic.

3.4 The Universalist Approach to AGI

A school of AGI research that doesn't fit neatly into any of the three categories reviewed above (symbolic, emergentist, hybrid) is what I call the "universalist approach". In this approach, one starts with AGI algorithms that would yield incredibly powerful general intelligence if supplied with massively, unrealistically much computing power; and then one tries to "scale them down," via adapting them to work using feasible computational resources. Historically, the roots of this approach may be traced to Solomonoff's pioneering work on the theory of induction (Solomonoff, 1964a,b).

The paradigm case of a universalist AGI approach is Marcus Hutter's AIXI system, which is based on the following simple concepts:

- An AGI system is going to be controlled by some program
- Instead of trying to figure out the right program via human wizardry, we can just write a "meta-algorithm" to search program space, and automatically find the right program for making the AGI smart, and then use that program to operate the AGI
- We can then repeat this meta-algorithm over and over, as the AGI gains more data about the world, so it will always have the operating program that's best according to all its available data

Marcus Hutter (Hutter, 2005) has proved that the AIXI system, which works basically as described in the above list, would be maximally generally intelligent, if the latter is defined appropriately in terms of maximizing computable reward functions in computable environments. The catch is that AIXI requires infinite processing power. But there's another version, AIXI^{ll}, that requires only an infeasibly massive finite amount of computing power.

Juergen Schmidhuber's Goedel Machine (Schmidhuber, 2006) operates differently in detail, but the concept is similar. At each step of the way, it takes the action that it can prove, according to its axiom system and its perceptual data, will be the best way to achieve its goals. Like AIXI, this is uncomputable in the most direct formulation, and computable but probably intractable in its most straightforward simplified formulations.

These theoretical approaches suggest a research program of “scaling down from infinity”, and finding practical, scalable ways of achieving AGI using similar ideas. Some promising results have been obtained, using simplified program space search to solve various specialized problems (Veness et al., 2011). But whether this approach can be used for human-level AGI, with feasible resource usage, remains uncertain. It’s a gutsy strategy, setting aside particularities of the human mind and brain, and focusing on what’s viewed as the mathematical essence of general intelligence.

A caricature of some common attitudes for and against the program search approach to AGI would be:

- **For:** The case of AGI with massive computational resources is an idealized case of AGI, similar to assumptions like the frictionless plane in physics, or the large population size in evolutionary biology. Now that we’ve solved the AGI problem in this simplified special case, we can use the understanding we’ve gained to address more realistic cases. This way of proceeding is mathematically and intellectually rigorous, unlike the more ad hoc approaches typically taken in the field. And we’ve already shown we can scale down our theoretical approaches to handle various specialized problems.
- **Against:** The theoretical achievement of advanced general intelligence using infinitely or unrealistically much computational resources, is a mathematical game which is only minimally relevant to achieving AGI using realistic amounts of resources. In the real world, the simple “trick” of exhaustively searching program space until you find the best program for your purposes, won’t get you very far. Trying to “scale down” from this simple method to something realistic isn’t going to work well, because real-world general intelligence is based on various complex, overlapping architectural mechanisms that just aren’t relevant to the massive-computational-resources situation.

4. Structures Underlying Human-Like General Intelligence

AGI is a very broad pursuit, not tied to the creation of systems emulating human-type general intelligence. However, if one temporarily restricts attention to AGI systems intended to vaguely emulate human functionality, then one can make significantly more intellectual progress in certain interesting directions. For example, by piecing together insights from the various architectures mentioned above, one can arrive at a rough idea regarding *what are the main aspects* that need to be addressed in creating a “human-level AGI” system.¹⁴

I will present here my rough understanding of the key aspects of human-level AGI in a series of seven figures, each adapted from a figure used to describe (all or part of) one of the AGI approaches listed above. The collective of these seven figures I will call the “integrative diagram.” When the term “architecture” is used in the context of these figures, it refers to an abstract cognitive architecture that may be realized in hardware, software, wetware or perhaps some other way. This “integrative diagram” is not intended as a grand theoretical conclusion, but rather as a didactic overview of the key elements involved in human-level general intelligence, expressed in a way that is not extremely closely tied to any one AGI architecture or theory, but represents a fair approximation of the AGI field’s overall understanding (inasmuch as such a diverse field can be said to have a coherent “overall understanding”).

14. The material in this section is adapted from a portion of the article (Goertzel, Iklé, and Wigmore, 2012)

HIGH LEVEL MIND ARCHITECTURE

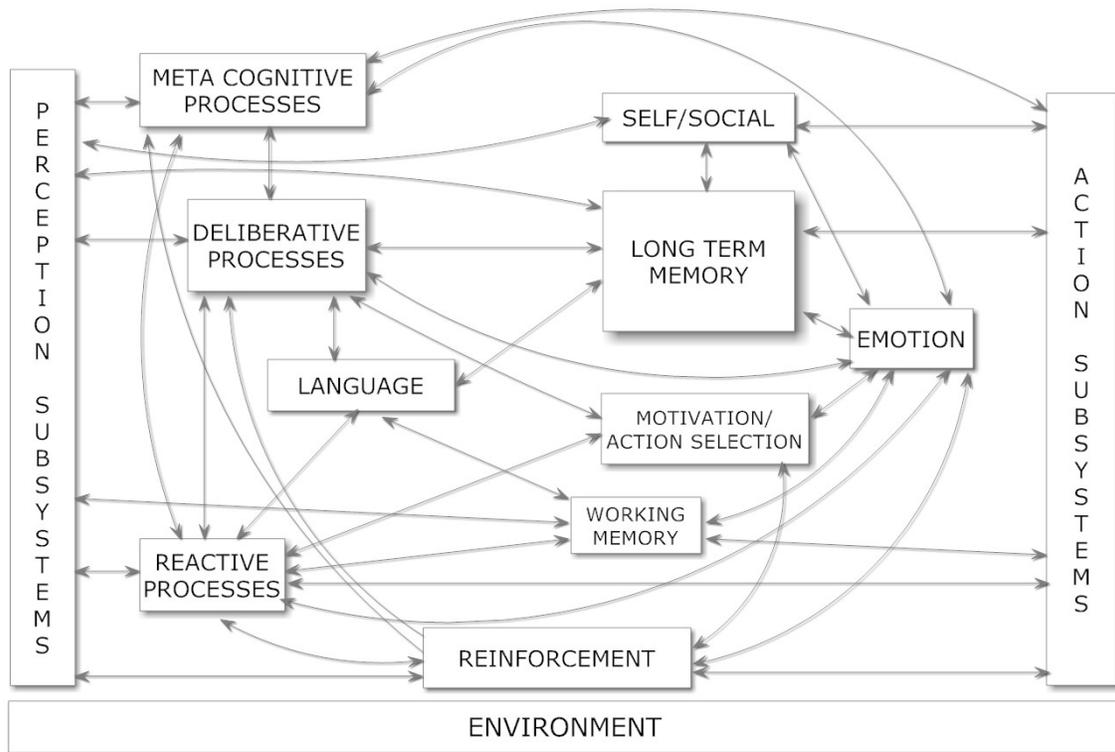


Figure 1: High-Level Structure of a Human-Like Mind

First, figure 1 gives a high-level breakdown of a human-like mind into components, based on Aaron Sloman's high-level cognitive-architectural sketch (Sloman, 2001). This diagram represents, roughly speaking, "modern common sense" about the architecture of a human-like mind. The separation between structures and processes, embodied in having separate boxes for Working Memory vs. Reactive Processes, and for Long Term Memory vs. Deliberative Processes, could be viewed as somewhat artificial, since in the human brain and most AGI architectures, memory and processing are closely integrated. However, the tradition in cognitive psychology is to separate out Working Memory and Long Term Memory from the cognitive processes acting thereupon, so I have adhered to that convention. The other changes from Sloman's diagram are the explicit inclusion of language, representing the hypothesis that language processing is handled in a somewhat special way in the human brain; and the inclusion of a reinforcement component parallel to the perception and action hierarchies, as inspired by intelligent control systems theory (e.g. Albus as mentioned above) and deep learning theory. Of course Sloman's high level diagram in its original form is intended as inclusive of language and reinforcement, but I felt it made sense to give them more emphasis.

Figure 2, modeling working memory and reactive processing, is essentially the LIDA diagram as given in prior papers by Stan Franklin, Bernard Baars and colleagues (Baars and Franklin, 2009;

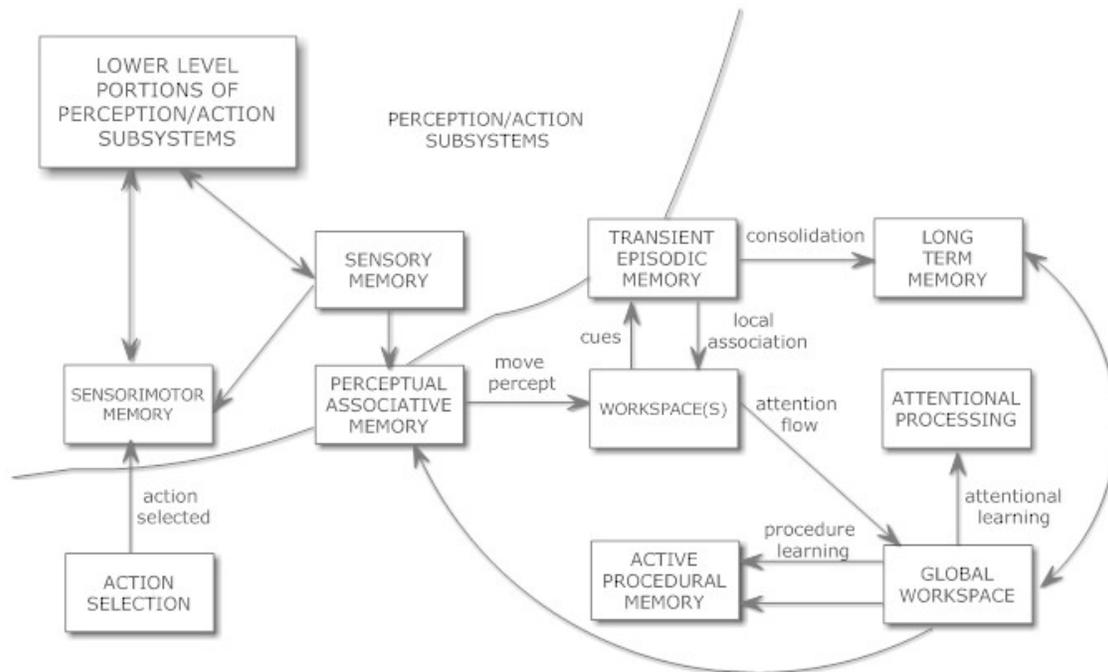


Figure 2: Architecture of Working Memory and Reactive Processing, closely modeled on the LIDA architecture

Franklin et al., 2012).¹⁵ The boxes in the upper left corner of the LIDA diagram pertain to sensory and motor processing, which LIDA does not handle in detail, and which are modeled more carefully by deep learning theory. The bottom left corner box refers to action selection, which in the integrative diagram is modeled in more detail by Psi. The top right corner box refers to Long-Term Memory, which the integrative diagram models in more detail as a synergetic multi-memory system (Figure 4).

Figure 3, modeling motivation and action selection, is a lightly modified version of the Psi diagram from Joscha Bach’s book *Principles of Synthetic Intelligence* (Bach, 2009). The main difference from Psi is that in the integrative diagram the Psi motivated action framework is embedded in a larger, more complex cognitive model. Psi comes with its own theory of working and long-term memory, which is related to but different from the one given in the integrative diagram – it views the multiple memory types distinguished in the integrative diagram as emergent from a common memory substrate. Psi comes with its own theory of perception and action, which seems broadly consistent with the deep learning approach incorporated in the integrative diagram. Psi’s handling of working memory lacks the detailed, explicit workflow of LIDA, though it seems broadly conceptually consistent with LIDA.

15. The original LIDA diagram refers to various “codelets”, a key concept in LIDA theory. I have replaced “attention codelets” here with “attention flow”, a more generic term. I suggest one can think of an attention codelet as a piece of information that it’s currently pertinent to pay attention to a certain collection of items together.

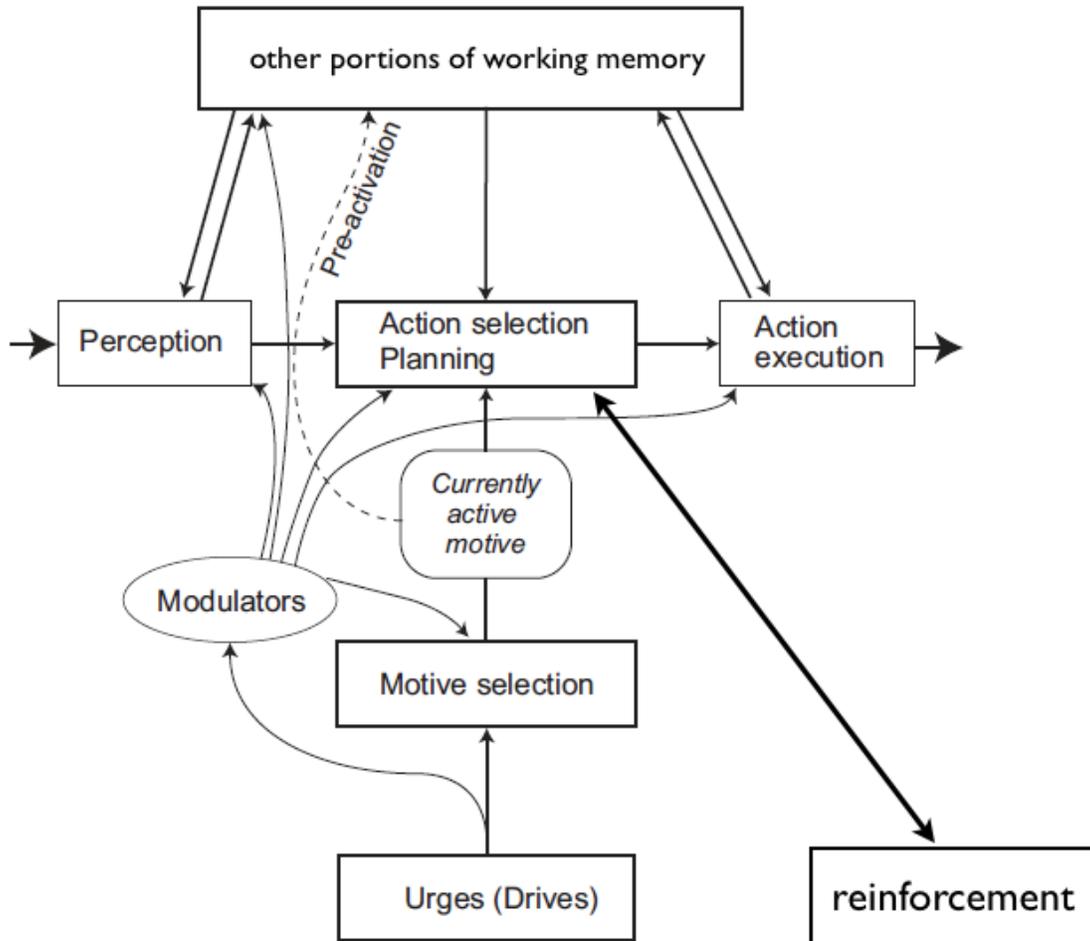


Figure 3: Architecture of Motivated Action

In Figure 3, the box labeled “Other parts of working memory” is labeled “Protocol and situation memory” in the original Psi diagram. The Perception, Action Execution and Action Selection boxes have fairly similar semantics to the similarly labeled boxes in the LIDA-like Figure 2, so that these diagrams may be viewed as overlapping. The LIDA model doesn’t explain action selection and planning in as much detail as Psi, so the Psi-like Figure 3 could be viewed as an elaboration of the action-selection portion of the LIDA-like Figure 2. In Psi, reinforcement is considered as part of the learning process involved in action selection and planning; in Figure 3 an explicit “reinforcement box” has been added to the original Psi diagram, to emphasize this.

Figure 4, modeling long-term memory and deliberative processing, is derived from my own prior work studying the “cognitive synergy” between different cognitive processes associated with different types of memory, and seeking to embody this synergy into the OpenCog system. The division into types of memory is fairly standard in the cognitive science field. Declarative, procedural, episodic and sensorimotor memory are routinely distinguished; we like to distinguish

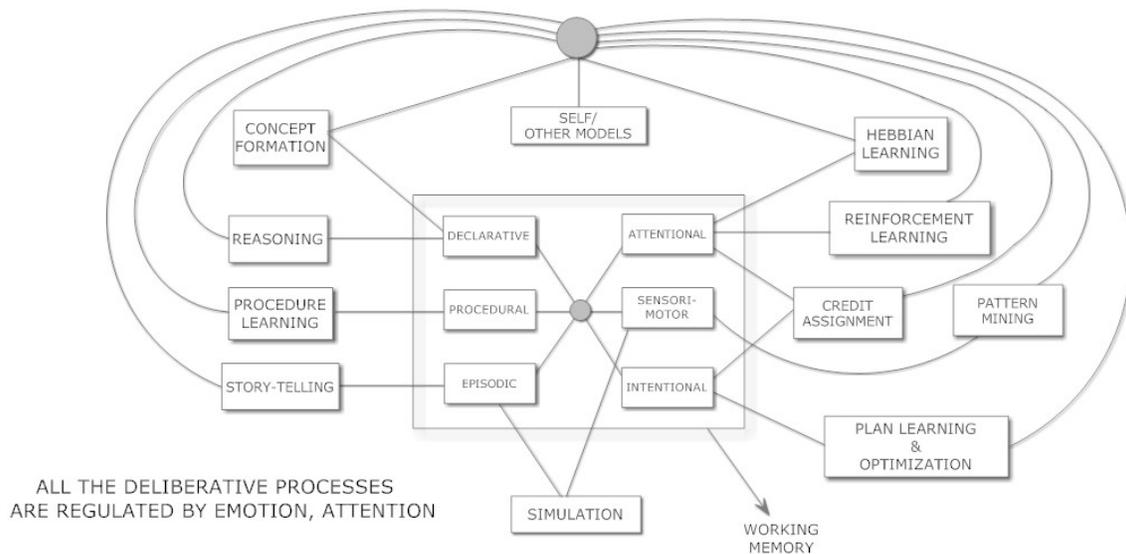


Figure 4: Architecture of Long-Term Memory and Deliberative and Metacognitive Thinking

attentional memory and intentional (goal) memory as well, and view these as the interface between long-term memory and the mind's global control systems. One focus of our AGI design work has been on designing learning algorithms, corresponding to these various types of memory, that interact with each other in a synergetic way (Goertzel, 2009), helping each other to overcome their intrinsic combinatorial explosions. There is significant evidence that these various types of long-term memory are differently implemented in the brain, but the degree of structure and dynamical commonality underlying these different implementations remains unclear (Gazzaniga, Ivry, and Mangun, 2009).

Each of these long-term memory types has its analogue in working memory as well. In some cognitive models, the working memory and long-term memory versions of a memory type and corresponding cognitive processes, are basically the same thing. OpenCog is mostly like this –it implements working memory as a subset of long-term memory consisting of items with particularly high importance values. The distinctive nature of working memory is enforced via using slightly different dynamical equations to update the importance values of items with importance above a certain threshold. On the other hand, many cognitive models treat working and long term memory as more distinct than this, and there is evidence for significant functional and anatomical distinctness in the brain in some cases. So for the purpose of the integrative diagram, it seemed best to leave working and long-term memory subcomponents as parallel but distinguished.

Figure 4 may be interpreted to encompass both workaday deliberative thinking and metacognition (“thinking about thinking”), under the hypothesis that in human beings and human-like minds, metacognitive thinking is carried out using basically the same processes as plain ordinary deliberative thinking, perhaps with various tweaks optimizing them for thinking about thinking. If it turns out that humans have, say, a special kind of reasoning faculty exclusively for metacognition, then the diagram would need to be modified. Modeling of self and others is understood to occur via a combination of metacognition and deliberative thinking, as well as via implicit adaptation based on reactive processing.

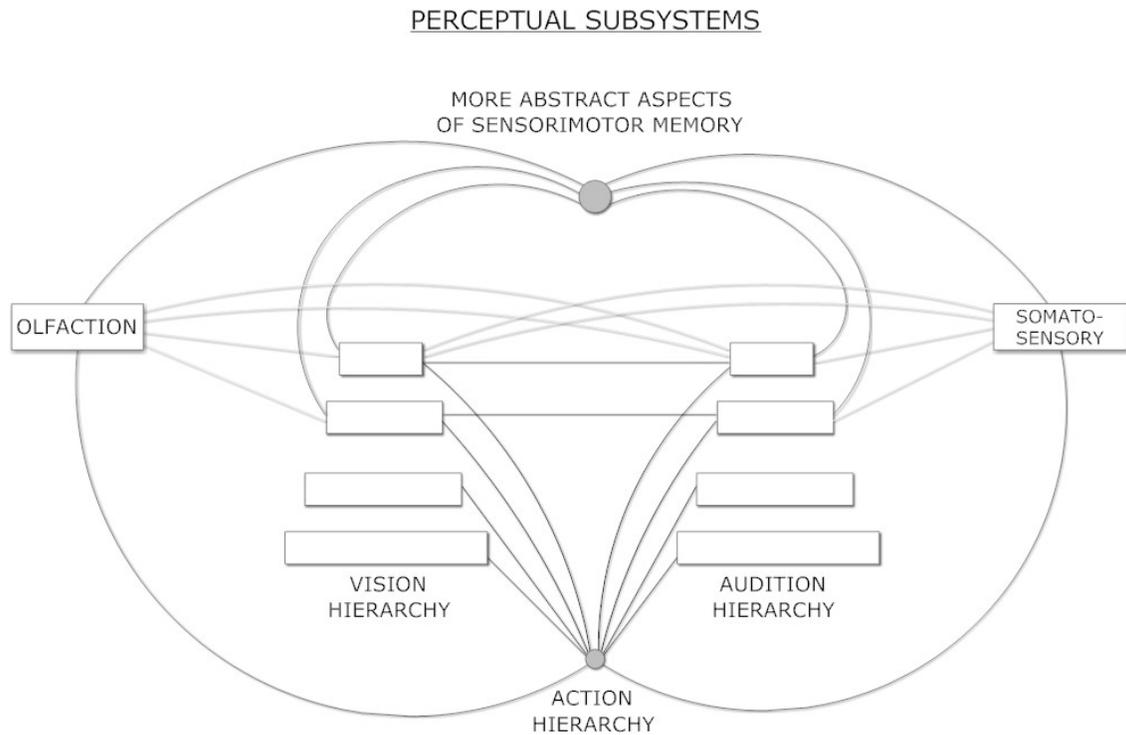


Figure 5: Architecture for Multimodal Perception

Figure 5 models perception, according to the concept of deep learning (Bengio, 2009; Anselmi et al., 2013). Vision and audition are modeled as deep learning hierarchies, with bottom-up and top-down dynamics. The lower layers in each hierarchy refer to more localized patterns recognized in, and abstracted from, sensory data. Output from these hierarchies to the rest of the mind is not just through the top layers, but via some sort of sampling from various layers, with a bias toward the top layers. The different hierarchies cross-connect, and are hence to an extent dynamically coupled together. It is also recognized that there are some sensory modalities that aren't strongly hierarchical, e.g touch and smell (the latter being better modeled as something like an asymmetric Hopfield net, prone to frequent chaotic dynamics (Li et al., 2005)) – these may also cross-connect with each other and with the more hierarchical perceptual subnetworks. Of course the suggested architecture could include any number of sensory modalities; the diagram is restricted to four just for simplicity.

The self-organized patterns in the upper layers of perceptual hierarchies may become quite complex and may develop advanced cognitive capabilities like episodic memory, reasoning, language learning, etc. A pure deep learning approach to intelligence argues that all the aspects of intelligence emerge from this kind of dynamics (among perceptual, action and reinforcement hierarchies). My own view is that the heterogeneity of human brain architecture argues against this perspective, and that deep learning systems are probably better as models of perception and action than of general cognition. However, the integrative diagram is not committed to my perspective on this – a deep-learning theorist could accept the integrative diagram, but argue that all the

other portions besides the perceptual, action and reinforcement hierarchies should be viewed as descriptions of phenomena that emerge in these hierarchies due to their interaction.

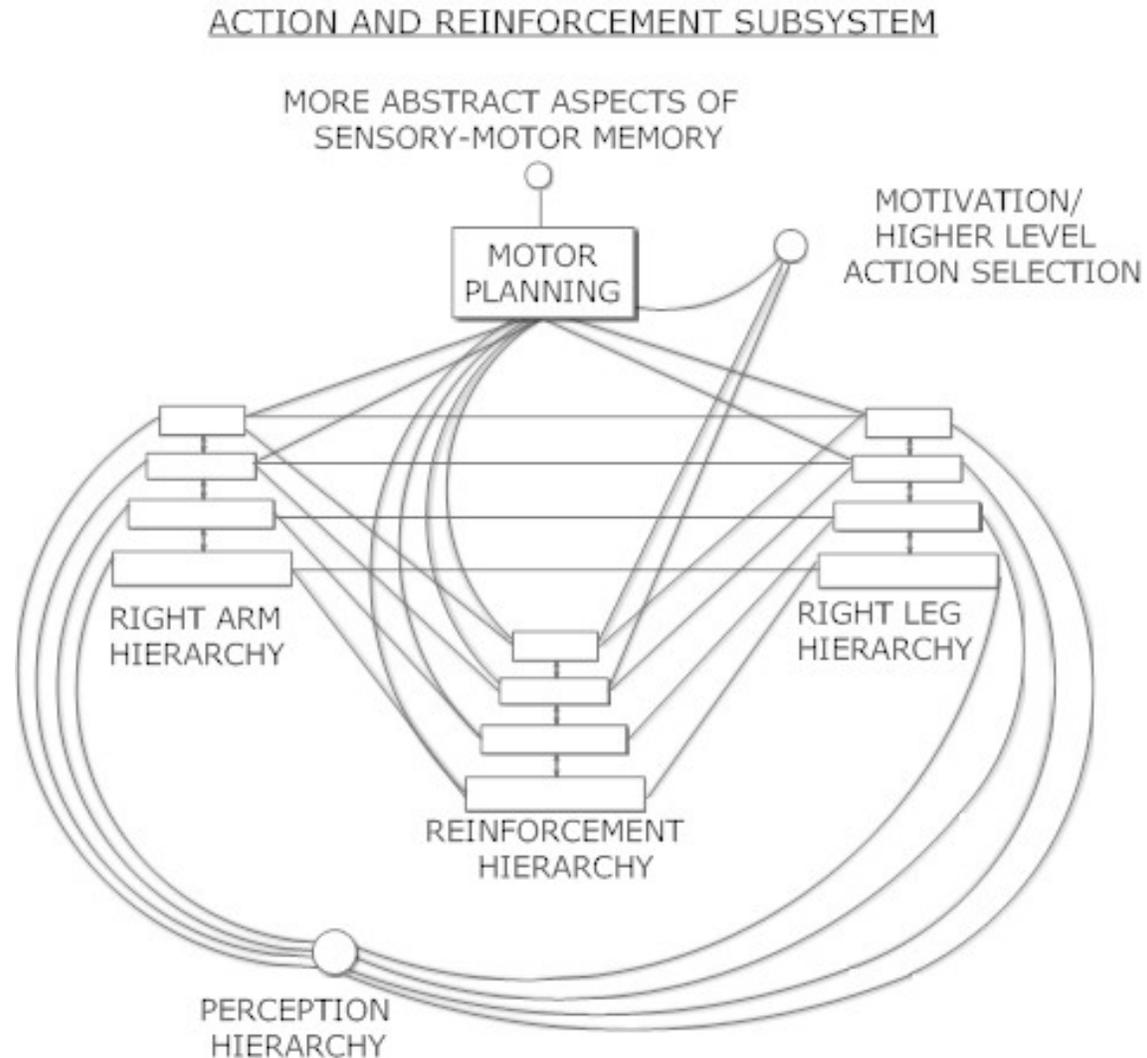


Figure 6: Architecture for Action and Reinforcement

Figure 6 shows an action subsystem and a reinforcement subsystem, parallel to the perception subsystem. Two action hierarchies, one for an arm and one for a leg, are shown for concreteness, but of course the architecture is intended to be extended more broadly. In the hierarchy corresponding to an arm, for example, the lowest level would contain control patterns corresponding to individual joints, the next level up to groupings of joints (like fingers), the next level up to larger parts of the arm (hand, elbow). The different hierarchies corresponding to different body parts cross-link, enabling coordination among body parts; and they also connect at multiple levels to perception hierarchies, enabling sensorimotor coordination. Finally there is a module for motor planning, which links tightly with all the motor hierarchies, and also overlaps with the more cognitive, inferential planning

activities of the mind, in a manner that is modeled different ways by different theorists. Albus (Albus, 2001) has elaborated this kind of hierarchy quite elaborately.

The reinforcement hierarchy in Figure 6 provides reinforcement to actions at various levels on the hierarchy, and includes dynamics for propagating information about reinforcement up and down the hierarchy.

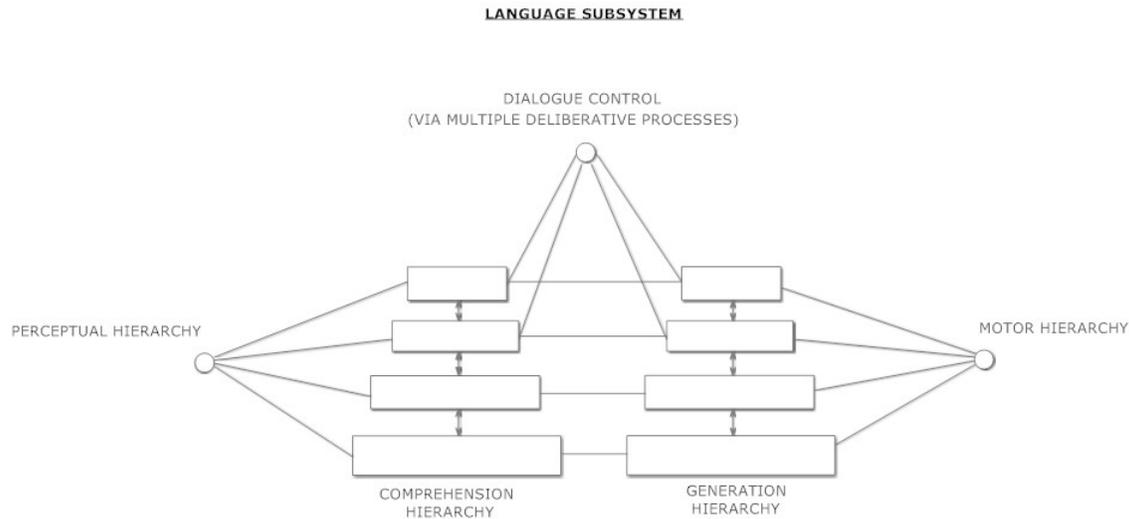


Figure 7: Architecture for Language Processing

Figure 7 deals with language, treating it as a special case of coupled perception and action. The traditional architecture of a computational language comprehension system is a pipeline (Jurafsky and James, 2000; Goertzel et al., 2010b), which is equivalent to a hierarchy with the lowest-level linguistic features (e.g. sounds, words) at the bottom, and the highest level features (semantic abstractions) at the top, and syntactic features in the middle. Feedback connections enable semantic and cognitive modulation of lower-level linguistic processing. Similarly, language generation is commonly modeled hierarchically, with the top levels being the ideas needing verbalization, and the bottom level corresponding to the actual sentence produced. In generation the primary flow is top-down, with bottom-up flow providing modulation of abstract concepts by linguistic surface forms.

This completes the posited, rough integrative architecture diagram for human-like general intelligence, split among 7 different pictures, formed by judiciously merging together architecture diagrams produced via a number of cognitive theorists with different, overlapping foci and research paradigms. One may wonder: Is anything critical left out of the diagram? A quick perusal of the table of contents of cognitive psychology textbooks suggests that if anything major is left out, it's also unknown to current cognitive psychology. However, one could certainly make an argument for explicit inclusion of certain other aspects of intelligence, that in the integrative diagram are left as implicit emergent phenomena. For instance, creativity is obviously very important to intelligence, but, there is no “creativity” box in any of these diagrams – because in our view, and the view of the cognitive theorists whose work we’ve directly drawn on here, creativity is best viewed as a process emergent from other processes that are explicitly included in the diagrams.

A high-level “cognitive architecture diagram” like this is certainly not a design for an AGI. Rather, it is more like a pointer in the direction of a requirements specification. These are, to a rough approximation, the aspects that must be taken into account, by anyone who wants to create a human-level AGI; and this is how these aspects appear to interact in the human mind. Different AGI approaches may account for these aspects and their interactions in different ways – e.g. via explicitly encoding them, or creating a system from which they can emerge, etc.

5. Metrics and Environments for Human-Level AGI

Science hinges on measurement; so if AGI is a scientific pursuit, it must be possible to measure what it means to achieve it.

Given the variety of approaches to AGI, it is hardly surprising that there are also multiple approaches to quantifying and measuring the achievement of AGI. However, things get a little simpler if one restricts attention to the subproblem of creating “human-level” AGI.

When one talks about AGI beyond the human level, or AGI that is very qualitatively different from human intelligence, then the measurement issue becomes very abstract – one basically has to choose a mathematical measure of general intelligence, and adopt it as a measure of success. This is a meaningful approach, yet also worrisome, because it’s difficult to tell, at this stage, what relation any of the existing mathematical measures of general intelligence is going to have to practical systems.

When one talks about human-level AGI, however, the measurement problem gets a lot more concrete: one can use tests designed to measure human performance, or tests designed relative to human behavior. The measurement issue then decomposes into two subproblems: quantifying achievement of the goal of human-level AGI, and measuring incremental progress toward that goal. The former subproblem turns out to be considerably more straightforward.

5.1 Metrics and Environments

The issue of metrics is closely tied up with the issue of “environments” for AGI systems. For AGI systems that are agents interacting with some environment, any method of measuring the general intelligence of these agents will involve the particulars of the AGI systems’ environments. If an AGI is implemented to control video game characters, then its intelligence must be measured in the video game context. If an AGI is built with solely a textual user interface, then its intelligence must be measured purely via conversation, without measuring, for example, visual pattern recognition.

And the importance of environments for AGI goes beyond the value of metrics. Even if one doesn’t care about quantitatively comparing two AGI systems, it may still be instructive to qualitatively observe the different ways they face similar situations in the same environment. Using multiple AGI systems in the same environment also increases the odds of code-sharing and concept-sharing between different systems. It makes it easier to conceptually compare what different systems are doing and how they’re working.

It is often useful to think in terms of “scenarios” for AGI systems, where a “scenario” means an environment plus a set of tasks defined in that environment, plus a set of metrics to measure performance on those tasks. At this stage, it is unrealistic to expect all AGI researchers to agree to conduct their research relative to the same scenario. The early-stage manifestations of different AGI approaches tend to fit naturally with different sorts of environments and tasks. However, to

whatever extent it is sensible for multiple AGI projects to share common environments or scenarios, this sort of cooperation should be avidly pursued.

5.2 Quantifying the Milestone of Human-Level AGI

A variety of metrics, relative to various different environments, may be used to measure achievement of the goal of “human-level AGI.” Examples include:

- the classic **Turing Test**, conceived as (roughly) “fooling a panel of college-educated human judges, during a one hour long interrogation, that one is a human being” (Turing, 1950) (and see (Hayes and Ford., 1995; French, 1996; Alvarado et al., 2002) for discussions of some of the test’s weaknesses).
- the **Virtual World Turing Test** occurring in an online virtual world, where the AGI and the human controls are controlling avatars (this is inclusive of the standard Turing Test if one assumes the avatars can use language) (Adams et al., 2012).
- Shane Legg’s **AIQ** measure Legg and Veness (2013), which is a computationally practical approximation to the algorithmic information theory based formalization of general intelligence given in (Legg and Hutter, 2007b). Work by Hernandez-Orallo and Dowe pursues a similar concept with different technical details(Hernández-Orallo and Dowe, 2010).
- Text compression – the idea being that any algorithm capable of understanding text should be transformable into an algorithm for compressing text based on the patterns it recognizes therein. This is the basis of the Hutter Prize (Hutter, 2006), a cash prize funded by Marcus Hutter which rewards data compression improvements on a specific 100 MB English text file, consisting of the first 100,000,000 characters of a certain version of English Wikipedia.
- the **Online University Student Test**, where an AGI 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.)(Adams et al., 2012).
- the **Robot University Student Test**, where an AGI has to obtain a college degree at an physical university, carrying out the same communications with the professors and the other students as a human student would, and also moving about the campus and handling relevant physical objects in a sufficient manner to complete the coursework(Adams et al., 2012).
- the **Artificial Scientist Test**, where an AGI 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 AGI has do original scientific research that wins a Nobel Prize)(Adams et al., 2012).

Each of these approaches has its pluses and minuses. None of them can sensibly be considered *necessary* conditions for human-level intelligence, but any of them may plausibly be considered sufficient conditions. The latter three have the disadvantage that they may not be achievable by every human – so they may set the bar a little too high. The former two have the disadvantage of requiring AGI systems to imitate humans, rather than just honestly being themselves; and it may

be that accurately imitating humans when one does not have a human body or experience, requires significantly *greater* than human level intelligence.

Regardless of the practical shortcomings of the above measures, though, I believe they are basically adequate as precisiations of “what it means to achieve human-level general intelligence.”

5.3 Measuring Incremental Progress Toward Human-Level AGI

While postulating criteria for assessing achievement of full human-level general intelligence seems relatively straightforward, positing good tests for *intermediate progress* toward the goal of human-level AGI seems much more difficult.

That is: it is not clear how to effectively measure whether one is, say, 50 percent of the way to human-level AGI? Or, say, 75 or 25 percent?

What I have found via a long series of discussions on this topic with a variety of AGI researchers is that:

- It’s possible to pose many “practical tests” of incremental progress toward human-level AGI, with the property that *if a proto-AGI system passes the test using a certain sort of architecture and/or dynamics, then this implies a certain amount of progress toward human-level AGI based on particular theoretical assumptions about AGI.*
- However, in each case of such a practical test, it seems intuitively likely *to a significant percentage of AGI researchers* that there is some way to “game” the test via designing a system specifically oriented toward passing that test, and which doesn’t constitute dramatic progress toward AGI.

A series of practical tests of this nature were discussed and developed at a 2009 gathering at the University of Tennessee, Knoxville, called the “AGI Roadmap Workshop,” which led to an article in AI Magazine titled *Mapping the Landscape of Artificial General Intelligence* (Adams et al., 2012). Among the tests discussed there were:

- The Wozniak “coffee test”¹⁶: go into an average American house and figure out how to make coffee, including identifying the coffee machine, figuring out what the buttons do, finding the coffee in the cabinet, etc.
- Story understanding – reading a story, or watching it on video, and then answering questions about what happened (including questions at various levels of abstraction)
- Passing the elementary school reading curriculum (which involves reading and answering questions about some picture books as well as purely textual ones)
- Learning to play an arbitrary video game based on experience only, or based on experience plus reading instructions
- Passing child psychologists’ typical evaluations aimed at judging whether a human preschool student is normally intellectually capable

16. The Wozniak coffee test, suggested by J. Storrs Hall, is so named due to a remark by Apple co-founder Steve Wozniak, to the effect that no robot will ever be able to go into a random American house and make a cup of coffee

One thing we found at the AGI Roadmap Workshop was that each of these tests seems to *some* AGI researchers to encapsulate the crux of the AGI problem, and to be unsolvable by any system not far along the path to human-level AGI – yet seems to other AGI researchers, with different conceptual perspectives, to be something probably game-able by narrow-AI methods. And of course, given the current state of science, there’s no way to tell which of these practical tests really can be solved via a narrow-AI approach, except by having a lot of researchers and engineers try really hard over a long period of time.

5.3.1 METRICS ASSESSING GENERALITY OF MACHINE LEARNING CAPABILITY

Complementing the above tests that are heavily inspired by human everyday life, there are also some more computer science oriented evaluation paradigms aimed at assessing AI systems going beyond specific tasks. For instance, there is a literature on multitask learning, where the goal for an AI is to learn one task quicker given another task solved previously (Thrun and Mitchell, 1995; Ben-David and Schuller, 2003; Taylor, Kuhlmann, and Stone, 2008). There is a literature on shaping, where the idea is to build up the capability of an AI by training it on progressively more difficult versions of the same tasks (Laud and Dejong, 2003; Li, Walsh, and Littman, 2006). Also, Achler (Achler, 2012a) has proposed criteria measuring the “flexibility of recognition” and posited this as a key measure of progress toward AGI.

While we applaud the work done in these areas, we also note it is an open question whether exploring these sorts of processes using mathematical abstractions, or in the domain of various machine-learning or robotics test problems, is capable of adequately addressing the problem of AGI. The potential problem with this kind of approach is that generalization among tasks, or from simpler to more difficult versions of the same task, is a process whose nature may depend strongly on the overall nature of the set of tasks and task-versions involved. Real-world humanly-relevant tasks have a subtlety of interconnectedness and developmental course that is not captured in current mathematical learning frameworks nor standard AI test problems.

To put it a little differently, it is possible that all of the following hold:

- the universe of real-world human tasks may possess a host of “special statistical properties that have implications regarding what sorts of AI programs will be most suitable
- exploring and formalizing and generalizing these statistical properties is an important research area; however,
- an easier and more reliable approach to AGI testing is to create a testing environment that embodies these properties implicitly, via constituting an emulation of the most cognitively meaningful aspects of the real-world human learning environment

Another way to think about these issues is to contrast the above-mentioned “AGI Roadmap Workshop” ideas with the “General Game Player (GGP) AI competition, in which AIs seek to learn to play games based on formal descriptions of the rules¹⁷. Clearly doing GGP well requires powerful AGI; and doing GGP even mediocly probably requires robust multitask learning and shaping. But it is unclear whether GGP constitutes a good approach to testing early-stage AI programs aimed at roughly humanlike intelligence. This is because, unlike the tasks involved in,

17. <http://games.stanford.edu/>

say, making coffee in an arbitrary house, or succeeding in preschool or university, the tasks involved in doing simple instances of GGP seem to have little relationship to humanlike intelligence or real-world human tasks.

So, an important open question is whether the class of statistical biases present in the set of real-world human environments tasks, has some sort of generalizable relevance to AGI beyond the scope of human-like general intelligence, or is informative only about the particularities of human-like intelligence. Currently we seem to lack any solid, broadly accepted theoretical framework for resolving this sort of question.

5.3.2 WHY IS MEASURING INCREMENTAL PROGRESS TOWARD AGI SO HARD?

A question raised by these various observations is whether there is some *fundamental reason* why it's hard to make an objective, theory-independent measure of intermediate progress toward advanced AGI, which respects the environment and task biased nature of human intelligence as well as the mathematical generality of the AGI concept. Is it just that we haven't been smart enough to figure out the right test – or is there some conceptual reason why the very notion of such a test is problematic?

Why might a solid, objective empirical test for intermediate progress toward humanly meaningful AGI be such a difficult project? One possible reason could be the phenomenon of “cognitive synergy” briefly noted above. In this hypothesis, for instance, it might be that there are 10 critical components required for a human-level AGI system. Having all 10 of them in place results in human-level AGI, but having only 8 of them in place results in having a dramatically impaired system – and maybe having only 6 or 7 of them in place results in a system that can hardly do anything at all.

Of course, the reality is not as strict as the simplified example in the above paragraph suggests. No AGI theorist has really posited a list of 10 crisply-defined subsystems and claimed them necessary and sufficient for AGI. We suspect there are many different routes to AGI, involving integration of different sorts of subsystems. However, if the cognitive synergy hypothesis is correct, then human-level AGI behaves *roughly* like the simplistic example in the prior paragraph suggests. Perhaps instead of using the 10 components, you could achieve human-level AGI with 7 components, but having only 5 of these 7 would yield drastically impaired functionality – etc. To mathematically formalize the cognitive synergy hypothesis becomes complex, but here we're only aiming for a qualitative argument. So for illustrative purposes, we'll stick with the “10 components” example, just for communicative simplicity.

Next, let's additionally suppose that for any given task, there are ways to achieve this task using a system that is much simpler than any subset of size 6 drawn from the set of 10 components needed for human-level AGI, but works much better for the task than this subset of 6 components (assuming the latter are used as a set of only 6 components, without the other 4 components).

Note that this additional supposition is a good bit stronger than mere cognitive synergy. For lack of a better name, I have called this hypothesis “tricky cognitive synergy” (Goertzel and Wigmore, 2011). Tricky cognitive synergy would be the case if, for example, the following possibilities were true:

- creating components to serve as parts of a synergetic AGI is *harder* than creating components intended to serve as parts of simpler AI systems without synergetic dynamics

- components capable of serving as parts of a synergetic AGI are necessarily *more complicated* than components intended to serve as parts of simpler AI systems

These certainly seem reasonable possibilities, since to serve as a component of a synergetic AGI system, a component must have the internal flexibility to usefully handle interactions with a lot of other components as well as to solve the problems that come its way.

If tricky cognitive synergy holds up as a property of human-level general intelligence, the difficulty of formulating tests for intermediate progress toward human-level AGI follows as a consequence. Because, according to the tricky cognitive synergy hypothesis, any test is going to be more easily solved by some simpler narrow AI process than by a *partially complete* human-level AGI system.

At the current stage in the development of AGI, we don't really know how big a role "tricky cognitive synergy" plays in the general intelligence. Quite possibly, 5 or 10 years from now someone will have developed wonderfully precise and practical metrics for the evaluation of incremental progress toward human-level AGI. However, it's worth carefully considering the possibility that fundamental obstacles, tied to the nature of general intelligence, stand in the way of this possibility.

6. What Would a General Theory of General Intelligence Look Like?

While most approaches to creating AGI are theoretically motivated in one way or another, nobody would claim there currently exists a thorough and systematic theory of AGI in the same sense that there exist theories of say, sorting algorithms, respiration, genetics, or near-equilibrium thermodynamics. Current AGI theory is a patchwork of overlapping concepts, frameworks and hypotheses, often synergetic and sometimes mutually contradictory. Current AGI system designs are usually *inspired by* theories, but do not have all their particulars *derived from* theories.

The creation of an adequate theoretical foundation for AGI is far beyond the scope of this review paper; however, it does seem worthwhile to briefly comment on what we may hope to get out of such a theory once it has been developed. Or in other words: What might a general theory of general intelligence look like?

Some of the things AGI researchers would like to do with a general theory of general intelligence are:

- Given a description of a set of goals and environments (and perhaps a probability distribution over these), and a set of computational resource restrictions, determine what is the system architecture that will display the maximum general intelligence relative to these goals and environments, subject to the given restrictions
- Given a description of a system architecture, figure out what are the goals and environments, with respect to which it will reach a relatively high level of general intelligence
- Given an intelligent system architecture, determine what sort of subjective experience the system will likely report having, in various contexts
- Given a set of subjective experiences and associated environments, determine what sort of intelligent system will likely have those experiences in those environments
- Find a practical way to synthesize a general-intelligence test appropriate for a given class of reasonably similar intelligent systems

- Identify the implicit representations of abstract concepts, arising within emergentist, hybrid, program learning based or other non-wholly-symbolic intelligent systems
- Given a certain intelligent system in a certain environment, predict the likely course of development of that system as it learns, experiences and grows
- Given a set of behavioral constraints (for instance, ethical constraints), estimate the odds that a given system will obey the constraints given certain assumptions about its environment. Determine architectures that, consistent with given computational resource constraints, provide an optimal balance between general intelligence for specified goals and environments, and adherence to given behavioral constraints
- What are the key structures and dynamics required for an AGI system to achieve human-level, human-like general intelligence within feasible computational resources?
- Predict the consequences of releasing an AGI into the world, depending on its level of intelligence and some specificities of its design
- Determine methods of assessing the ethical character of an AGI system, both in its current form and in future incarnations likely to develop from its current form (for discussion of various issues regarding the ethics of advanced AGI see (Goertzel and Pitt, 2012; Bostrom, 2014; Hibbard, 2012; Yudkowsky, 2008))

Anyone familiar with the current state of AGI research will find it hard to suppress a smile at this ambitious list of objectives. At the moment we would seem very far from having a theoretical understanding capable of thoroughly addressing any of these points, in a practically useful way. It is unclear to how far the limits of mathematics and computing will allow us to progress toward theoretical goals such as these. However: the further we can get in this direction, the better off the AGI field will be.

At the moment, AGI system design is as much artistic as scientific, relying heavily on the designer's scientific intuition. AGI implementation and testing are interwoven with (more or less) inspired tinkering, according to which systems are progressively improved internally as their behaviors are observed in various situations. This sort of approach is not unworkable, and many great inventions have been created via similar processes. It's unclear how necessary or useful a more advanced AGI theory will be for the creation of practical AGI systems. But it seems likely that, the further we can get toward a theory providing tools to address questions like those listed above, the more systematic and scientific the AGI design process will become, and the more capable the resulting systems.

It's possible that a thorough, rigorous theory of AGI will emerge from the mind of some genius AGI researcher, in one fell swoop – or from the mind of one of the early AGI successes itself! However, it appears more probable that the emergence of such a theory will be a gradual process, in which theoretical and experimental developments progress hand in hand.

7. Conclusion

Given the state of the art in AGI research today, what can we say about the core AGI hypothesis? Is it actually the case that creating generally intelligent systems, requires fundamentally different

concepts and approaches than creating more specialized, “narrow AI” systems? Is there a deep necessity for considering “AGI” as its own distinctive pursuit?

Personally I am confident the answer to this question is “yes.” However, setting aside intuition and looking only at the available relevant science and engineering results, I would have to say that the jury is still out. The narrow AI approach has not led to dramatic progress toward AGI goals; but at the present time, the AGI achievements of researchers explicitly working toward AGI (myself included) have also been relatively modest. There exist a number of theoretical frameworks explaining why AGI is profoundly distinct from narrow AI; but none of these frameworks can be considered thoroughly empirically validated.

The next question, then, is: What is being done – and what should be done – to further explore the core AGI hypothesis, and move toward its verification or falsification? It seems that to move the AGI field rapidly forward, one of the two following things must happen:

- The emergence, within the AGI community, of a broadly accepted theory of general intelligence – including a characterization of what it is, and a theory of what sorts of architecture can be expected to work for achieving human-level AGI using realistic computational resources; or
- The demonstration of an AGI system that qualitatively appears, to both novice and expert observers, to demonstrate a dramatic and considerable amount of general intelligence. For instance: a robot that can do a variety of preschool-type activities in a flexible and adaptive way; or a chatbot that can hold an hour’s conversation without sounding insane or resorting to repeating catch-phrases, etc.

Neither of these occurrences would rigorously prove the core AGI hypothesis. However, either of them would build confidence in the core AGI hypothesis: in the first case because there would be a coherent and broadly accepted theory implying the core AGI hypothesis; in the second case because we would have a practical demonstration that an AGI perspective *has in fact worked better* for creating AGI than a narrow AI approach.

These are still early days for AGI; and yet, given the reality of exponential technological advance (Kurzweil, 2005), this doesn’t necessarily imply that dramatic success is a long way off. There is a real possibility of dramatic, interlinked progress in AGI design, engineering, evaluation and theory in the relatively near future – in the next few decades, and potentially even the next few years. No one can accurately predict the course of development of any research area; but it’s interesting that in a survey of researchers at the AGI-2010 conference, the majority of respondents felt that human-level AGI was likely to arise before 2050, and some were much more optimistic (Seth Baum and Goertzel, 2011). Optimism regarding the near advent of advanced AGI is controversial, but is a position held by an increasing plurality of the AGI community, who are working hard to make their hopes and projections rapidly eventuate.

References

Achler, T. 2012a. Artificial General Intelligence Begins with Recognition: Evaluating the Flexibility of Recognition. In *Theoretical Foundations of Artificial General Intelligence*. Springer. 197–217.

- Achler, T. 2012b. Towards Bridging the Gap Between Pattern Recognition and Symbolic Representation Within Neural Networks. *Workshop on Neural-Symbolic Learning and Reasoning, AAAI-2012*.
- Adams, S.; Arel, I.; Bach, J.; Coop, R.; Furlan, R.; Goertzel, B.; Hall, J. S.; Samsonovich, A.; Scheutz, M.; Schlesinger, M.; et al. 2012. Mapping the landscape of human-level artificial general intelligence. *AI Magazine* 33(1):25–42.
- Albus, J. S. 2001. *Engineering of mind: An introduction to the science of intelligent systems*. Wiley.
- Alvarado, N.; Adams, S. S.; Burbeck, S.; and Latta, C. 2002. Beyond the Turing test: Performance metrics for evaluating a computer simulation of the human mind. In *The 2nd International Conference on Development and Learning*, 147–152. IEEE.
- Anderson, J. R., and Lebiere, C. 2003. The Newell test for a theory of cognition. *Behavioral and Brain Sciences* 26(05):587–601.
- Anselmi, F.; Leibo, J. Z.; Rosasco, L.; Mutch, J.; Tacchetti, A.; and Poggio, T. 2013. Magic Materials: a theory of deep hierarchical architectures for learning sensory representations.
- Arel, I.; Rose, D.; and Coop, R. 2009. Destin: A scalable deep learning architecture with application to high-dimensional robust pattern recognition. In *Proc. AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*, 1150–1157.
- Arel, I.; Rose, D.; and Karnowski, T. 2009. A deep learning architecture comprising homogeneous cortical circuits for scalable spatiotemporal pattern inference. In *NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications*.
- Baars, B. J., and Franklin, S. 2009. Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness* 1(01):23–32.
- Bach, J. 2009. *Principles of synthetic intelligence PSI: an architecture of motivated cognition*, volume 4. Oxford University Press.
- Baranès, A., and Oudeyer, P.-Y. 2009. R-IAC: Robust intrinsically motivated exploration and active learning. *Autonomous Mental Development, IEEE Transactions on* 1(3):155–169.
- Ben-David, S., and Schuller, R. 2003. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*. Springer. 567–580.
- Bengio, Y. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning* 2(1):1–127.
- Binet, A., and Simon, T. 1916. *The development of intelligence in children: The Binet-Simon Scale*. Number 11. Williams & Wilkins Company.
- Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Brooks, R. A. 2002. *Flesh and machines: How robots will change us*. Pantheon Books New York.

- Cassimatis, N. 2007. Adaptive algorithmic hybrids for human-level Artificial Intelligence. In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, 94–112.
- Damer, B.; Newman, P.; Gordon, R.; and Barbalet, T. 2010. The EvoGrid: simulating pre-biotic emergent complexity.
- De Garis, H.; Shuo, C.; Goertzel, B.; and Ruiting, L. 2010. A world survey of artificial brain projects, Part I: Large-scale brain simulations. *Neurocomputing* 74(1):3–29.
- Duch, W.; Oentaryo, R. J.; and Pasquier, M. 2008. Cognitive Architectures: Where do we go from here? In *Proceedings of the First Conference on Artificial General Intelligence*, volume 171, 122–136.
- Dye, L. 2010. Are Dolphins Also Persons? *ABC News, Feb. 24 2010*.
- Franklin, S., and Graesser, A. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent agents III: agent theories, architectures, and languages*. Springer. 21–35.
- Franklin, S.; Strain, S.; Snaider, J.; McCall, R.; and Faghihi, U. 2012. Global workspace theory, its LIDA model and the underlying neuroscience. *Biologically Inspired Cognitive Architectures* 1:32–43.
- French, R. M. 1996. Subcognition and the Limits of the Turing Test. *Machines and thought* 11–26.
- Frye, J.; Ananthanarayanan, R.; and Modha, D. S. 2007. Towards real-time, mouse-scale cortical simulations. *CoSyNe: Computational and Systems Neuroscience, Salt Lake City, Utah*.
- Gardner, H. 1999. *Intelligence reframed: Multiple intelligences for the 21st century*. Basic Books.
- Gazzaniga, M. S.; Ivry, R. B.; and Mangun, G. R. 2009. *Cognitive Neuroscience: The Biology of the Mind*. W W Norton.
- Goertzel, B., and Pennachin, C. 2007. *Artificial General Intelligence*. Springer.
- Goertzel, B., and Pitt, J. 2012. Nine Ways to Bias Open-Source AGI Toward Friendliness. *Journal of Evolution and Technology* 22:1.
- Goertzel, B., and Wigmore, J. 2011. Cognitive Synergy Is Tricky. *Chinese Journal of Mind and Computation*.
- Goertzel, B.; Lian, R.; Arel, I.; de Garis, H.; and Chen, S. 2010a. A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. *Neurocomputing* 74(1):30–49.
- Goertzel, B.; Pennachin, C.; Araujo, S.; Silva, F.; Queiroz, M.; Lian, R.; Silva, W.; Ross, M.; Vepstas, L.; and Senna, A. 2010b. A general intelligence oriented architecture for embodied natural language processing. In *3d Conference on Artificial General Intelligence (AGI-2010)*. Atlantis Press.
- Goertzel, B.; Pitt, J.; Wigmore, J.; Geisweiller, N.; Cai, Z.; Lian, R.; Huang, D.; and Yu, G. 2011. Cognitive Synergy between Procedural and Declarative Learning in the Control of Animated and Robotic Agents Using the OpenCogPrime AGI Architecture. In *Proceedings of AAAI-11*.

- Goertzel, B.; Iklé, M.; and Wigmore, J. 2012. The Architecture of Human-Like General Intelligence. In *Theoretical Foundations of Artificial General Intelligence*. Springer. 123–144.
- Goertzel, B. 2009. OpenCogPrime: A cognitive synergy based architecture for artificial general intelligence. In *Proceedings of ICCI'09: 8th IEEE International Conference on Cognitive Informatics*, 60–68. IEEE.
- Goertzel, B. 2010. Toward a formal characterization of real-world general intelligence. In *Proceedings of the Third Conference on Artificial General Intelligence*, 19–24.
- Goertzel, B. 2014. Artificial General Intelligence. *Japanese Artificial Intelligence Society Magazine*, 2014-1.
- Gregory, R. J. 2004. *Psychological testing: History, principles, and applications*. Allyn & Bacon.
- Gubrud, M. A. 1997. Nanotechnology and international security. In *Fifth Foresight Conference on Molecular Nanotechnology*, 1.
- Hammer, B., and Hitzler, P. 2007. *Perspectives of neural-symbolic integration*, volume 77. Springer.
- Han, J.; Zeng, S.; Tham, K.; Badgero, M.; and Weng, J. 2002. Dav: A humanoid robot platform for autonomous mental development. In *Development and Learning, 2002. Proceedings. The 2nd International Conference on*, 73–81. IEEE.
- Hawkins, J., and Blakeslee, S. 2007. *On intelligence*. Macmillan.
- Hayes, P., and Ford., K. 1995. Turing Test Considered Harmful. *IJCAI-14*.
- Hernández-Orallo, J., and Dowe, D. L. 2010. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence* 174(18):1508–1539.
- Hibbard, B. 2012. Avoiding unintended AI behaviors. In *Artificial General Intelligence*. Springer. 107–116.
- Horwitz, B.; Friston, K. J.; and Taylor, J. G. 2000. Neural modeling and functional brain imaging: an overview. *Neural networks* 13(8):829–846.
- Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer.
- Hutter, M. 2006. Human Knowledge Compression Contest. <http://prize.hutter1.net/>.
- Izhikevich, E. M., and Edelman, G. M. 2008. Large-scale model of mammalian thalamocortical systems. *Proc. of the national academy of sciences* 105(9):3593–3593.
- Jilk, D. J., and Lebiere, C. 2008. SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence* 20:197–218.
- Jurafsky, D., and James, H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech.

- Just, M. A., and Varma, S. 2007. The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, and Behavioral Neuroscience* 7:153–191.
- Kaplan, F. 2008. Neurorobotics: an experimental science of embodiment. *Frontiers in neuroscience* 2(1):22.
- Koza, J. R. 1992. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press.
- Krichmar, J. L., and Edelman, G. M. 2006. Principles underlying the construction of brain-based devices. In *Proceedings of AISB*, volume 6, 37–42.
- Kurzweil, R. 2005. *The singularity is near: When humans transcend biology*. Penguin.
- Laird, J. E.; Wray, R.; Marinier, R.; and Langley, P. 2009. Claims and challenges in evaluating human-level intelligent systems. In *Proceedings of the Second Conference on Artificial General Intelligence*, 91–96.
- Laird, J. 2012. *The Soar cognitive architecture*. MIT Press.
- Langley, P. 2005. An adaptive architecture for physical agents. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 18–25. IEEE.
- Laud, A., and Dejong, G. 2003. The influence of reward on the speed of reinforcement learning. *Proc. of the 20th International Conf. on Machine Learning*.
- Le, Q. V. 2013. Building high-level features using large scale unsupervised learning. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8595–8598. IEEE.
- Legg, S., and Hutter, M. 2007a. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications* 157:17.
- Legg, S., and Hutter, M. 2007b. Universal intelligence: A definition of machine intelligence. *Minds and Machines* 17(4):391–444.
- Legg, S., and Veness, J. 2013. An approximation of the universal intelligence measure. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*. Springer. 236–249.
- Lenat, D. B., and Guha, R. V. 1989. *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc.
- Li, G.; Lou, Z.; Wang, L.; Li, X.; and Freeman, W. J. 2005. Application of chaotic neural model based on olfactory system on pattern recognitions. In *Advances in Natural Computation*. Springer. 378–381.
- Li, L.; Walsh, T.; and Littman, M. 2006. Towards a unified theory of state abstraction for MDPs. *Proc. of the ninth international symposium on AI and mathematics*.

- Markram, H. 2006. The blue brain project. *Nature Reviews Neuroscience* 7(2):153–160.
- Metta, G.; Sandini, G.; Vernon, D.; Natale, L.; and Nori, F. 2008. The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, 50–56. ACM.
- Modayil, J., and Kuipers, B. 2007. Autonomous development of a grounded object ontology by a learning robot. In *Proceedings of the national conference on Artificial intelligence*, volume 22, 1095. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
- Mugan, J., and Kuipers, B. 2008. Towards the application of reinforcement learning to undirected developmental learning. *International Conf. on Epigenetic Robotics*.
- Mugan, J., and Kuipers, B. 2009. Autonomously Learning an Action Hierarchy Using a Learned Qualitative State Representation. In *IJCAI*, 1175–1180.
- Muggleton, S. 1991. Inductive logic programming. *New generation computing* 8(4):295–318.
- Nestor, A., and Kokinov, B. 2004. Towards Active Vision in the DUAL Cognitive Architecture. *International Journal on Information Theories and Applications* 11.
- Nilsson, N. J. 2005. Human-level artificial intelligence? Be serious! *AI magazine* 26(4):68.
- Nilsson, N. J. 2007. The physical symbol system hypothesis: status and prospects. In *50 years of artificial intelligence*. Springer. 9–17.
- Oudeyer, P.-Y., and Kaplan, F. 2006. Discovering communication. *Connection Science* 18(2):189–206.
- Pfeifer, R., and Bongard, J. 2007. *How the body shapes the way we think: a new view of intelligence*. MIT press.
- Reeke Jr, G. N.; Sporns, O.; and Edelman, G. M. 1990. Synthetic neural modeling: the Darwin's series of recognition automata. *Proceedings of the IEEE* 78(9):1498–1530.
- Richardson, M., and Domingos, P. 2006. Markov logic networks. *Machine learning* 62(1-2):107–136.
- Rosbe, J.; Chong, R. S.; and Kieras, D. E. 2001. Modeling with Perceptual and Memory Constraints: An EPIC-Soar Model of a Simplified Enroute Air Traffic Control Task. *SOAR Technology Inc. Report*.
- Russell, S. J., and Norvig, P. 2010. *Artificial intelligence: a modern approach*. Prentice Hall.
- Samsonovich, A. V. 2010. Toward a Unified Catalog of Implemented Cognitive Architectures. *BICA* 221:195–244.
- Schmidhuber, J. 1991a. Curious model-building control systems.. *Proc. International Joint Conf. on Neural Networks*.

- Schmidhuber, J. 1991b. A possibility for implementing curiosity and boredom in model-building neural controllers. *Proc. of the International Conf. on Simulation of Adaptive Behavior: From Animals to Animats*.
- Schmidhuber, J. 1995. Reinforcement-driven information acquisition in non-deterministic environments. *Proc. ICANN'95*.
- Schmidhuber, J. 2003. Exploring the predictable. In *Advances in evolutionary computing*. Springer. 579–612.
- Schmidhuber, J. 2006. Godel machines: Fully Self-Referential Optimal Universal Self-Improvers. In Goertzel, B., and Pennachin, C., eds., *Artificial General Intelligence*. 119–226.
- Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3(03):417–424.
- Seth Baum, B. G., and Goertzel, T. 2011. Technological Forecasting and Social Change. *Technological Forecasting and Social Change*.
- Shapiro, S. C.; Rapaport, W. J.; Kandeler, M.; Johnson, F. L.; and Goldfain, A. 2007. Metacognition in SNePS. *AI Magazine* 28(1):17.
- Shastri, L., and Ajjanagadde, V. 1993. From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and brain sciences* 16(3):417–451.
- Silver, R.; Boahen, K.; Grillner, S.; Kopell, N.; and Olsen, K. L. 2007. Neurotech for neuroscience: unifying concepts, organizing principles, and emerging tools. *The Journal of Neuroscience* 27(44):11807–11819.
- Sloman, A. 2001. Varieties of affect and the cogaff architecture schema. In *Proceedings of the AISBOI symposium on emotions, cognition, and affective computing. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour*.
- Socher, R.; Huval, B.; Bath, B. P.; Manning, C. D.; and Ng, A. Y. 2012. Convolutional-Recursive Deep Learning for 3D Object Classification. In *NIPS*, 665–673.
- Solomonoff, R. J. 1964a. A formal theory of inductive inference. Part I. *Information and control* 7(1):1–22.
- Solomonoff, R. J. 1964b. A formal theory of inductive inference. Part II. *Information and control* 7(2):224–254.
- Spearman, C. 1904. General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology* 15(2):201–292.
- Sun, R., and Zhang, X. 2004. Top-down versus bottom-up learning in cognitive skill acquisition. *Cognitive Systems Research* 5(1):63–89.
- Taylor, M. E.; Kuhlmann, G.; and Stone, P. 2008. Transfer Learning and Intelligence: an Argument and Approach. *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS* 171:326.

- Terman, L. M. 1915. The mental hygiene of exceptional children. *The Pedagogical Seminary* 22(4):529–537.
- Thrun, S., and Mitchell, T. 1995. Lifelong robot learning. *Robotics and Autonomous Systems*.
- Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 433–460.
- Veness, J.; Ng, K. S.; Hutter, M.; Uther, W.; and Silver, D. 2011. A monte-carlo aixo approximation. *Journal of Artificial Intelligence Research* 40(1):95–142.
- Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence*. Springer.
- Wang, P. 2009. Embodiment: Does a Laptop Have a Body? In *Proceedings of AGI-09*, 74–179.
- Weng, J., and Hwang, W.-S. 2006. From neural networks to the brain: Autonomous mental development. *Computational Intelligence Magazine, IEEE* 1(3):15–31.
- Weng, J.; Hwang, W. S.; Zhang, Y.; Yang, C.; and Smith, R. 2000. Developmental humanoids: Humanoids that develop skills automatically. In *Proc. The First IEEE-RAS International Conference on Humanoid Robots*, 7–8. Citeseer.
- Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks*. Oxford University Press. 303.