Chapter 1

The Architecture of Human-Like General Intelligence

Ben Goertzel[1], Matt Ikle' [2] and Jared Wigmore[3]
[1] *Novamente LLC,*
1405 Bernerd Place, Rockville MD
ben@goertzel.org
[2] *Dept. of Mathematics and Computing, Adams State College, Alamoso CO*
[3] *School of Design, Hong Kong Polytechnic University*

By exploring the relationships between different AGI architectures, one can work toward a holistic cognitive model of human-level intelligence. In this vein, here an integrative architecture diagram for human-like general intelligence is proposed, via merging of lightly modified version of prior diagrams including Aaron Sloman's high-level cognitive model, Stan Franklin and the LIDA group's model of working memory and the cognitive cycle, Joscha Bach and Dietrich Dorner's Psi model of motivated action and cognition, James Albus's three-hierarchy intelligent robotics model, and the author's prior work on cognitive synergy in deliberative thought and metacognition, along with ideas from deep learning and computational linguistics. The purpose is not to propose an actual merger of the various AGI systems considered, but rather to highlight the points of compatibility between the different approaches, as well as the differences of both focus and substance. The result is perhaps the most comprehensive architecture diagram of human-cognition yet produced, tying together all key aspects of human intelligence in a coherent way that is not tightly bound to any particular cognitive or AGI theory. Finally, the question of the dynamics associated with the architecture is considered, including the potential that human-level intelligence requires cognitive synergy between these various components is considered; and the possibility of a "trickiness" property causing the intelligence of the overall system to be badly suboptimal if any of the components are missing or insufficiently cooperative. One idea emerging from these dynamic consideration is that implementing the *whole* integrative architecture diagram may be necessary for achieving anywhere near human-level, human-like general intelligence.

1.1. Introduction

Cognitive science appears to have a problem with integrative understanding. Over the last few decades, cognitive science has discovered a great deal about the structures and dynamics underlying the human mind. However, as in many other branches of science, there has been more focus on detailed analysis of individual aspects, than on unified holistic understanding. As a result of this tendency, there are not many compelling examples of holistic "cognitive architecture diagrams" for human intelligence – diagrams systematically

laying out all the pieces needed to generate human intelligence and how they interact with each other.

Part of the reason why global human cognitive architecture diagrams are not so common is, of course, a lack of agreement in the field regarding all the relevant issues. Since there are multiple opinions regarding nearly every aspect of human intelligence, it would be difficult to get two cognitive scientists to fully agree on every aspect of an overall human cognitive architecture diagram. Prior attempts to outline detailed mind architectures have tended to follow highly specific theories of intelligence, and hence have attracted only moderate interest from researchers not adhering to those theories. An example is Minsky's work presented in The Emotion Machine,[?] which arguably does constitute an architecture diagram for the human mind, but which is only loosely grounded in current empirical knowledge and stands more as a representation of Minsky's own intuitive understanding.

On the other hand, AGI appears to have a problem with the mutual comprehensibility and comparability of different research approaches. The AGI field has in recent years seen a proliferation of cognitive architectures, each purporting to cover all aspects needed for the creation of human-level general intelligence. However, the differences of language and focus among the various approaches has often made it difficult for researchers to fully understand each others' work, let alone collaborate effectively.

This chapter describes a conceptual experiment aimed at addressing both of the above problems together We aim to present a coherent, overall architecture diagram for human, and human-like, general intelligence, via combining the architecture diagrams associated with a number of contemporary AGI architectures. While the exercise is phrased in terms of diagrams, of course the underlying impetus is conceptual integration; and our hope is that the exercise described here will serve as a starting point for ongoing exploration of the relationships between multiple AGI architectures and cognitive theories.

The architecture diagram we give here does not reflect our own idiosyncratic understanding of human intelligence, as much as a combination of understandings previously presented by multiple researchers (including ourselves), arranged according to our own taste in a manner we find conceptually coherent. With this in mind, we call it "the integrative diagram" (a longer, grander and more explanatory name would be "The First Integrative Human-Like Cognitive Architecture Diagram"). We have made an effort to ensure that as many pieces of the integrative diagram as possible are well grounded in psychological and even neuroscientific data, rather than mainly embodying speculative notions; however, given the current state of knowledge, this could not be done to a complete extent, and there is still some speculation involved here and there.

While based largely on understandings of human intelligence, the integrative diagram is intended to serve as an architectural outline for human-like general intelligence more broadly. For example, the OpenCog AGI architecture which we have co-created is explicitly not intended as a precise emulation of human intelligence, and does many things quite differently than the human mind, yet can still fairly straightforwardly be mapped into the integrative diagram.

Finally, having presented the integrative diagram which focuses on *structure*, we

present some comments on the dynamics corresponding to that structure, focusing on the notion of *cognitive synergy*: the hypothesis that multiple subsystems of a generally intelligent system, focused on learning regarding different sorts of information, must interact in such a way as to actively aid each other in overcoming combinatorial explosions. This represents a fairly strong hypothesis regarding how the different components in the integrative diagram interact with each other. Further, it gives a way of addressing one of the more vexing issues in the AGI field: the difficulty of measuring partial progress toward human-level AGI. We will conjecture that this difficulty is largely attributable to a "trickiness" property of cognitive synergy in the integrative diagram. One of the upshots of our discussion of dynamics is: We consider it likely that to achieve anything remotely like human-like general intelligence, it will be necessary to implement basically all the components in the integrative diagram, in a thorough and richly interconnected way. Implementing half the boxes in the diagram is not likely to get us to a system with half the general intelligence of a human. The architecture of human-like cognition is a richly interconnected whole.

1.2. Key Ingredients of the Integrative Human-Like Cognitive Architecture Diagram

In assembling the integrative diagram, we have drawn on the work of researchers at the intersection of AGI and cognitive science – that is, researchers largely motivated by human cognitive science and neuroscience, but with aims of producing comprehensive architectures for human-like AGI. The main ingredients used in assembling the diagram are:

- Aaron Sloman's high-level architecture diagram of human intelligence,[?] drawn from his CogAff archtiecture, which is strikes me as a particularly clear embodiment of "modern common sense" regarding the overall architecture of the human mind. We have added only a couple items to Sloman's high-level diagram, which we felt deserved an explicit high-level role that he did not give them: emotion, language and reinforcement.
- The LIDA architecture diagram presented by Stan Franklin and Bernard Baars.[?] We think LIDA is an excellent model of working memory and what Sloman calls "reactive processes", with well-researched grounding in the psychology and neuroscience literature. We have adapted the LIDA diagram only very slightly for use here, changing some of the terminology on the arrows, and indicating where parts of the LIDA diagram indicate processes elaborated in more detail elsewhere in the integrative diagram.
- The architecture diagram of the Psi model of motivated cognition, presented by Joscha Bach in[?] based on prior work by Dietrich Dorner.[?] This diagram is presented without significant modification; however it should be noted that Bach and Dorner present this diagram in the context of larger and richer cognitive models, the other aspects of which are not all incorporated in the integrative diagram.
- James Albus's three-hierarchy model of intelligence,[?] involving coupled perception, action and reinforcement hierarchies. Albus's model, utilized in the creation

of intelligent unmanned automated vehicles, is a crisp embodiment of many ideas emergent from the field of intelligent control systems.

- Deep learning networks as a model of perception (and action and reinforcement learning), as embodied for example in the work of Itamar Arel[?] and Jeff Hawkins.[?] The integrative diagram adopts this as the basic model of the perception and action subsystems of human intelligence. Language understanding and generation are also modeled according to this paradigm.
- The OpenCog[?] integrative AGI architecture (in which I have played a key role), which places greatest focus on various types of long-term memory and their interrelationship, and is used mainly to guide the integrative architecture's treatment of these matters.

Most of these ingredients could be interpreted as holistic explanations of human-like intelligence on their own. However, each of them places a focus i a different place, more elaborated in some regards than others. So it is interesting to collage the architecture diagrams from the different approaches together, and see what results. The product of this exercise does not accord precisely with any of the component AGI architectures, and is not proposed as an architecture diagram for an AGI. However, we believe it has value as an exercise in integrative cognitive science. It is a mind-architecture diagram, drawing preferentially on different cognitive-science-inspired AGI approaches in those aspects where they have been most thoroughly refined.

One possible negative reaction to the integrative diagram might be to say that it's a kind of Frankenstein monster, piecing together aspects of different theories in a way that violates the theoretical notions underlying all of them! For example, the integrative diagram takes LIDA as a model of working memory and reactive processing, but from the papers on LIDA it's unclear whether the creators of LIDA construe it more broadly than that. The deep learning community tends to believe that the architecture of current deep learning networks, in itself, is close to sufficient for human-level general intelligence – whereas the integrative diagram appropriates the ideas from this community mainly for handling perception, action and language. Etc.

On the other hand, in a more positive perspective, one could view the integrative diagram as consistent with LIDA, but merely providing much more detail on some of the boxes in the LIDA diagram (e.g. dealing with perception and long-term memory). And one could view the integrative diagram as consistent with consistent with the deep learning paradigm – via viewing it, not as a description of components to be explicitly implemented in an AGI system, but rather as a description of the key structures and processes that must emerge in deep learning network, based on its engagement with the world, in order for it to achieve human-like general intelligence.

It seems to us that different communities of cognitive science and AGI researchers have focused on different aspects of intelligence, and have thus each created models that are more fully fleshed out in some aspects than others. But these various models all link together fairly cleanly, which is not surprising as they are all grounded in the same data re-

garding human intelligence. Many judgment calls must be made in fusing multiple models in the way that the integrative diagram does, but we feel these can be made without violating the spirit of the component models. In assembling the integrative diagram, we have made these judgment calls as best we can, but we're well aware that different judgments would also be feasible and defensible. Revisions are likely as time goes on, not only due to new data about human intelligence but also to evolution of understanding regarding the best approach to model integration.

Another possible argument against the ideas presented here is that there's nothing new – all the ingredients presented have been given before elsewhere. To this our retort is to quote Pascal: "Let no one say that I have said nothing new ... the arrangement of the subject is new." The various architecture diagrams incorporated into the integrative diagram are either extremely high level (Sloman's diagram) or focus primarily on one aspect of intelligence, treating the others very concisely by summarizing large networks of distinction structures and processes in small boxes. The integrative diagram seeks to cover all aspects of human-like intelligence at a roughly equal granularity – a different arrangement.

1.3. An Architecture Diagram for Human-Like General Intelligence

The integrative diagram is presented here in a series of seven figures.

Figure 1.1 gives a high-level breakdown into components, based on Sloman's high-level cognitive-architectural sketch.[?] This diagram represents, roughly speaking, "modern common sense" about how a human-like mind is architected. 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 we 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 we felt it made sense to give them more emphasis.

Figure 1.2, modeling working memory and reactive processing, is essentially the LIDA diagram as given in prior papers by Stan Franklin, Bernard Baars and colleagues.[?] 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 1.4).

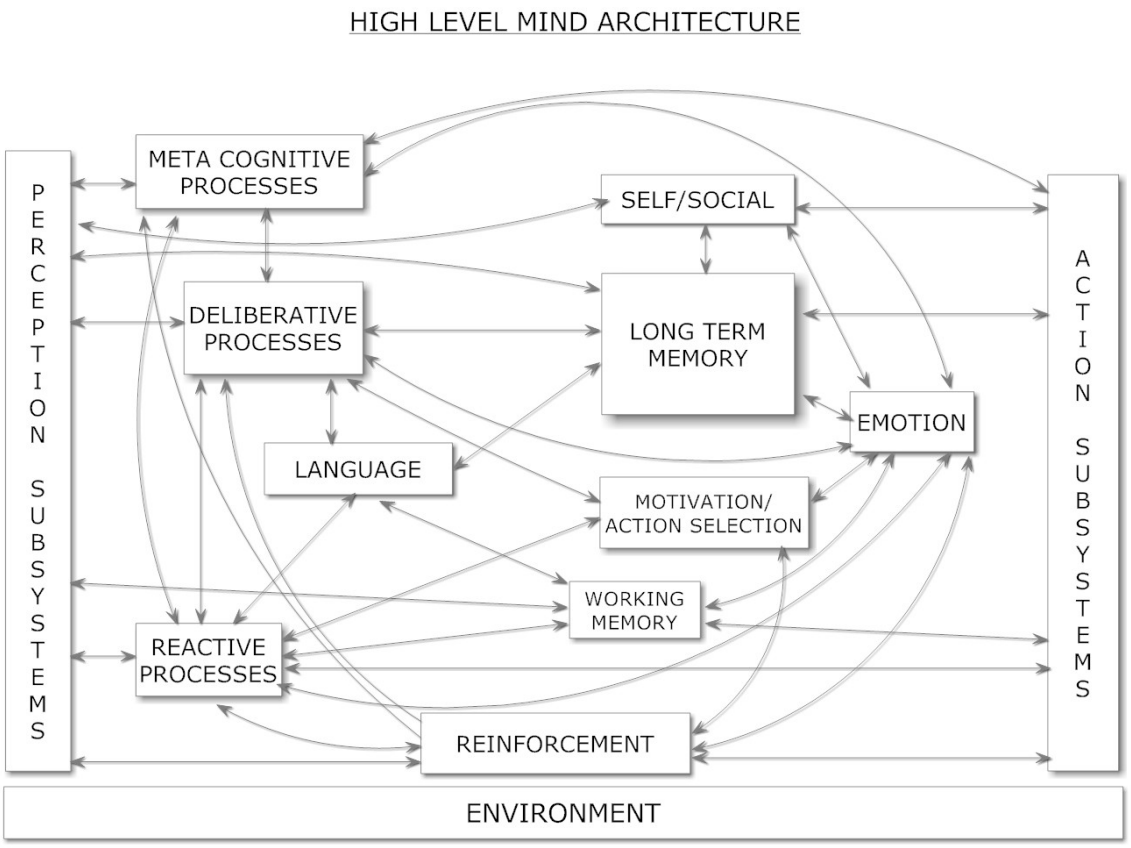

Fig. 1.1. High-Level Architecture of a Human-Like Mind

The original LIDA diagram refers to various "codelets", a key concept in LIDA theory. We have replaced "attention codelets" here with "attention flow", a more generic term. We 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 1.3, modeling motivation and action selection, is a lightly modified version of the Psi diagram from Joscha Bach's book *Principles of Synthetic Intelligence*.[?] 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.

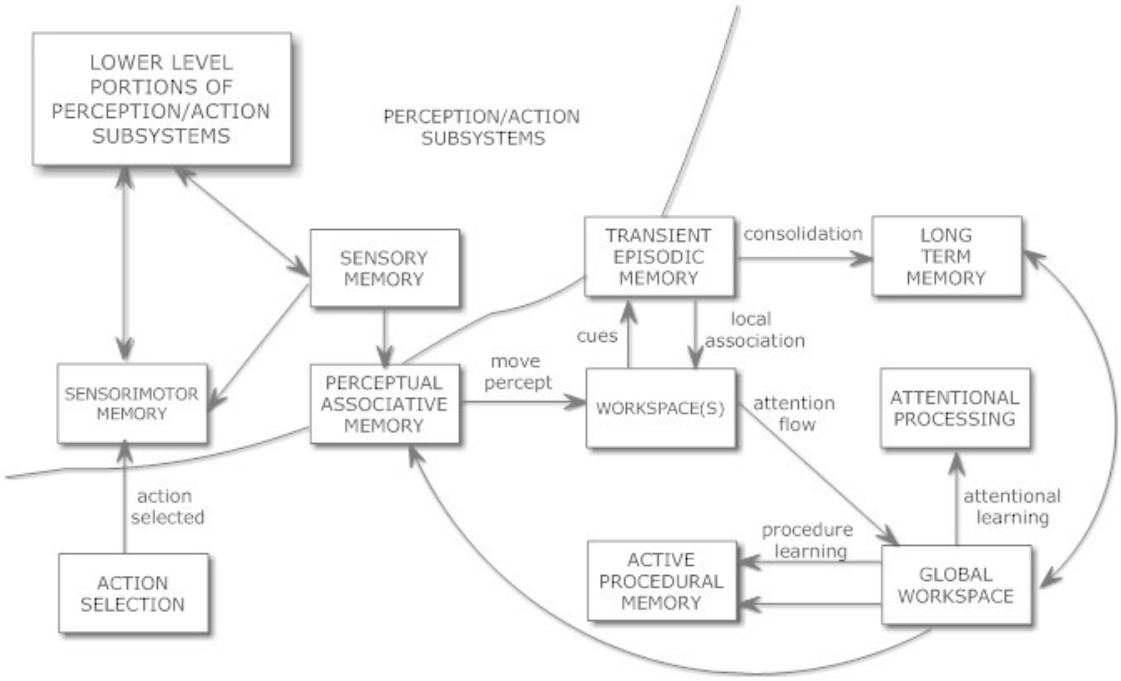

Fig. 1.2. Architecture of Working Memory and Reactive Processing, closely modeled on the LIDA architecture

In Figure 1.3, 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 1.2, 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 1.3 could be viewed as an elaboration of the action-selection portion of the LIDA-like Figure 1.2. In Psi, reinforcement is considered as part of the learning process involved in action selection and planning; in Figure 1.3 an explicit "reinforcement box" has been added to the original Psi diagram, to emphasize this.

Figure 1.4, modeling long-term memory and deliberative processing, is derived from our own prior work studying the "cognitive synergy" between different cognitive processes associated with different types of memory. The division into types of memory is fairly standard. Declarative, procedural, episodic and sensorimotor memory are routinely distinguished; we like to distinguish 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,[?] helping each other to overcome their intrinsic combinatorial explosions. There is significant evidence that these various types of long-term memory are differently imple-

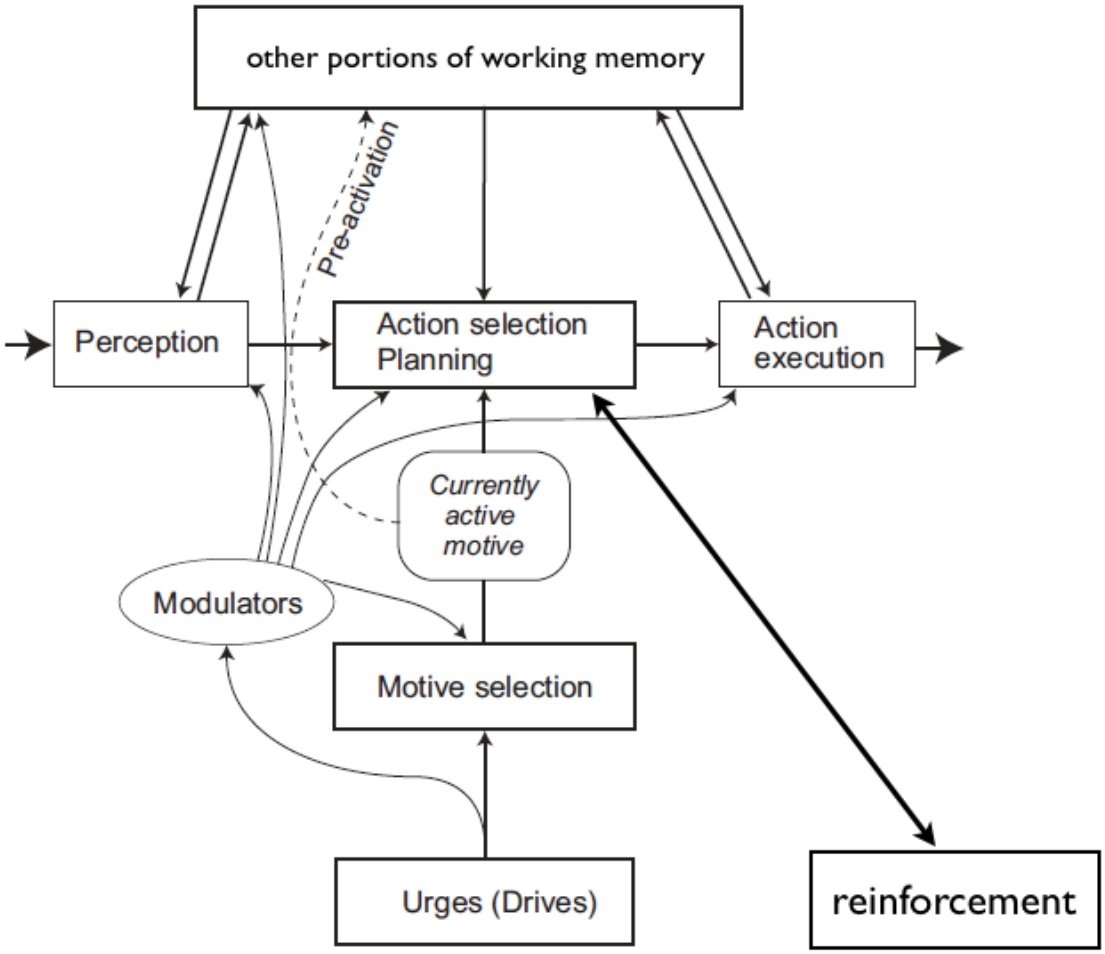

Fig. 1.3. Architecture of Motivated Action

mented in the brain, but the degree of structure and dynamical commonality underlying these different implementations remains unclear.

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

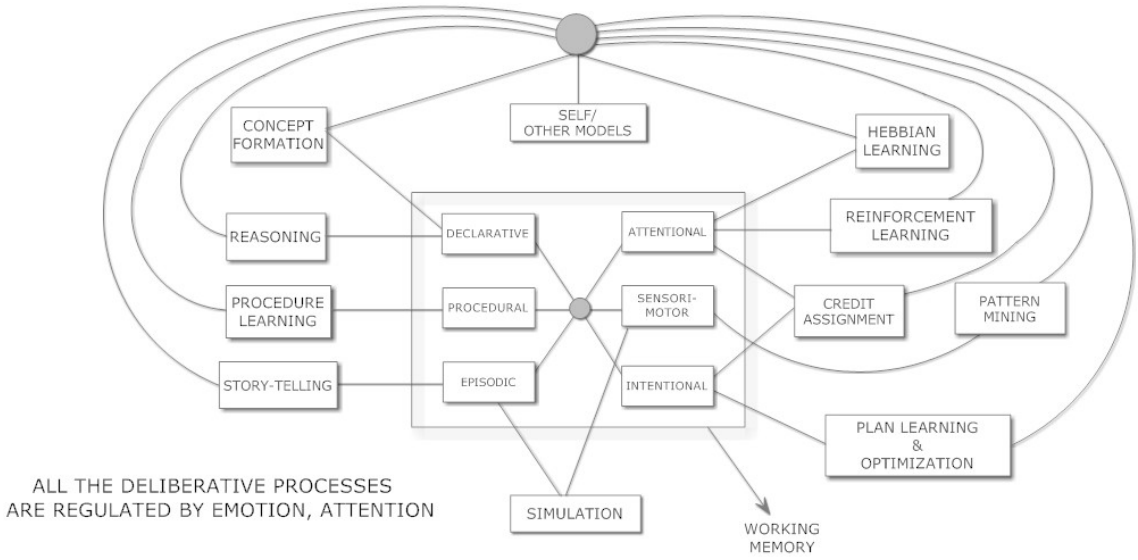

Fig. 1.4. Architecture of Long-Term Memory and Deliberative and Metacognitive Thinking

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 1.4 also encompasses metacognition, 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 1.5 models perception, according to the basic ideas of deep learning theory. 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**?**) – 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

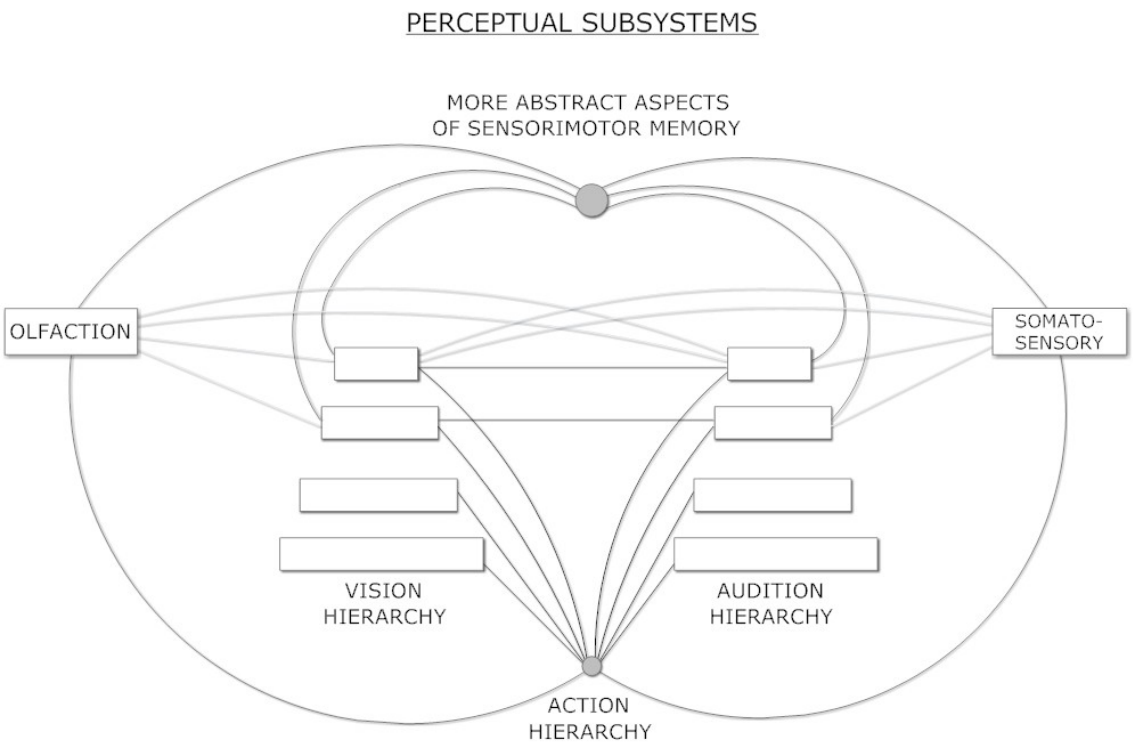

Fig. 1.5. Architecture for Multimodal Perception

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). Our 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 our 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 1.6 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

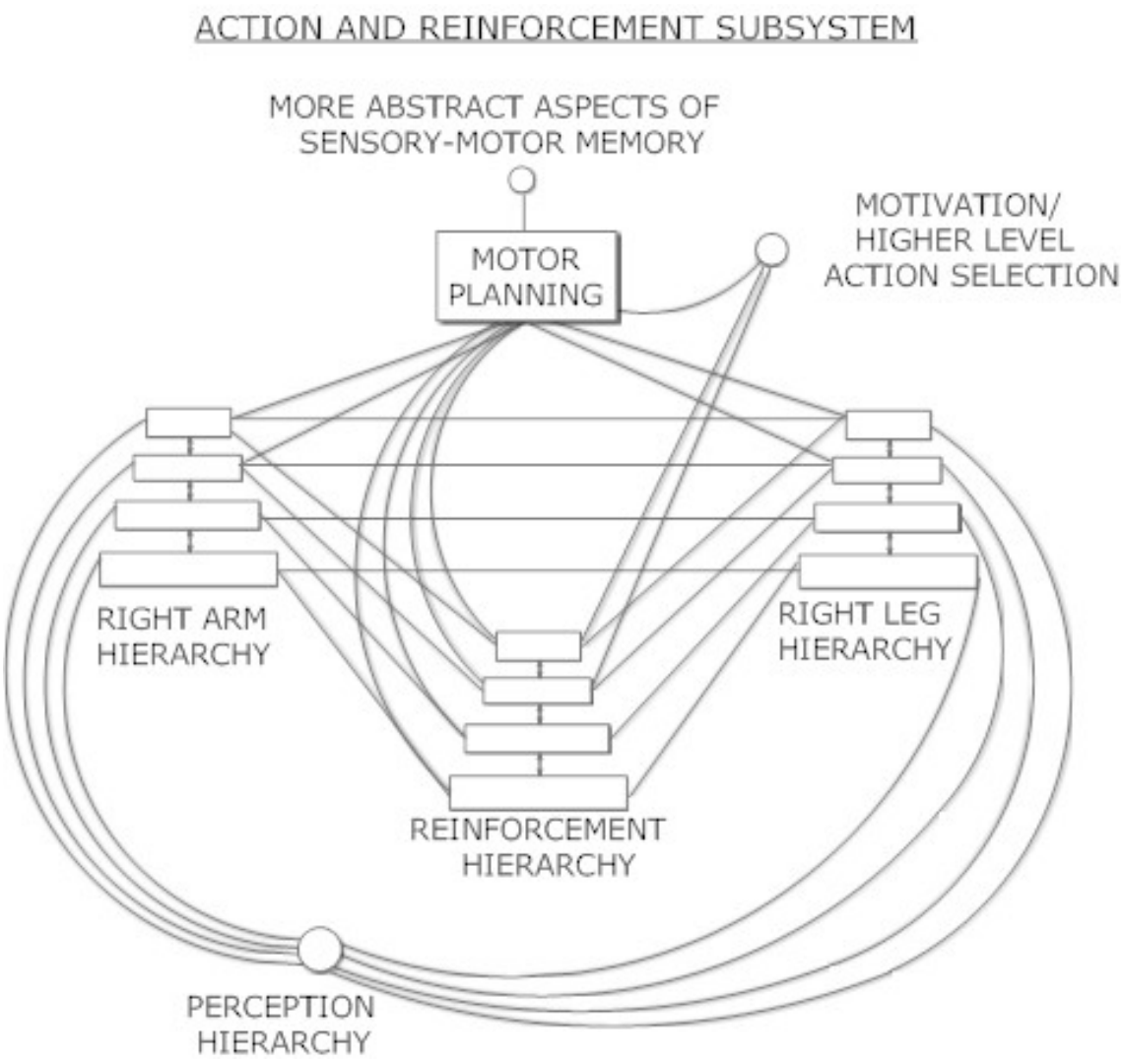

Fig. 1.6. Architecture for Action and Reinforcement

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[?] has elaborated this kind of hierarchy quite elaborately.

The reward hierarchy in Figure 1.6 provides reinforcement to actions at various levels on the hierarchy, and includes dynamics for propagating information about reinforcement up and down the hierarchy.

Figure 1.7 deals with language, treating it as a special case of coupled perception and

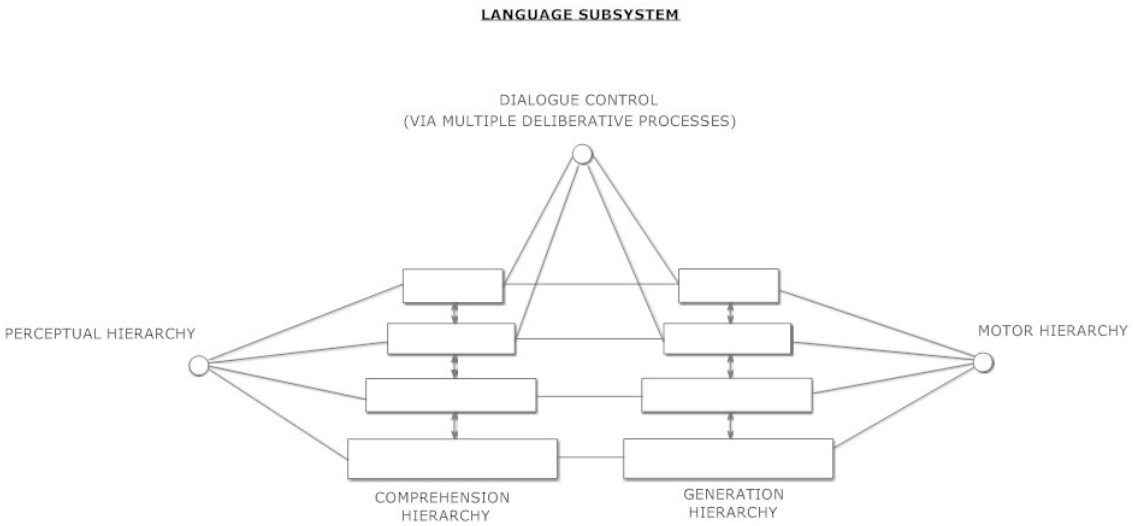

Fig. 1.7. Architecture for Language Processing

action. The traditional architecture of a computational language comprehension system is a pipeline[?,?] 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.

So, that's it – an 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.

Is anything critical left out of the diagram? A quick perusal of the table of contents of cognitive psychology textbooks suggests to me 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.

1.4. Interpretation and Application of the Integrative Diagram

A tongue-partly-in-cheek definition of a biological pathway is "a subnetwork of a biological network, that fits on a single journal page." Cognitive architecture diagrams have a similar property – they are crude abstractions of complex structures and dynamics, sculpted in accordance with the size of the printed page, and the tolerance of the human eye for absorbing diagrams, and the tolerance of the human author for making diagrams.

However, sometimes constraints – even arbitrary ones – are useful for guiding creative efforts, due to the fact that they force choices. Creating an architecture for human-like general intelligence that fits in a few (okay, 7) fairly compact diagrams, requires one to make many choices about what features and relationships are most essential. In constructing the integrative diagram, we have sought to make these choices, not purely according to our own tastes in cognitive theory or AGI system design, but according to a sort of blend of the taste and judgment of a number of scientists whose views we respect, and who seem to have fairly compatible, complementary perspectives.

What is the use of a cognitive architecture diagram like this? It can help to give newcomers to the field a basic idea about what is known and suspected about the nature of human-like general intelligence. Also, it could potentially be used as a tool for cross-correlating different AGI architectures. If everyone who authored an AGI architecture would explain how their architecture accounts for each of the structures and processes identified in the integrative diagram, this would give a means of relating the various AGI designs to each other.

The integrative diagram could also be used to help connect AGI and cognitive psychology to neuroscience in a more systematic way. In the case of LIDA, a fairly careful correspondence has been drawn up between the LIDA diagram nodes and links and various neural structures and processes.[?] Similar knowledge exists for the rest of the integrative diagram, though not organized in such a systematic fashion. A systematic curation of links between the nodes and links in the integrative diagram and current neuroscience knowledge, would constitute an interesting first approximation of the holistic cognitive behavior of the human brain.

Finally (and harking forward to the next section), the big omission in the integrative diagram is *dynamics*. Structure alone will only get you so far, and you could build an AGI system with reasonable-looking things in each of the integrative diagram's boxes, interrelating according to the given arrows, and yet still fail to make a viable AGI system. Given the limitations the real world places on computing resources, it's not enough to have adequate representations and algorithms in all the boxes, communicating together properly and capable doing the right things given sufficient resources. Rather, one needs to have all the boxes filled in properly with structures and processes that, when they act together using feasible computing resources, will yield appropriately intelligent behaviors via their cooperative activity. And this has to do with the complex interactive dynamics of all the processes in all the different boxes – which is something the integrative diagram doesn't touch at all. This brings us again to the network of ideas we've discussed under the name

of "cognitive synergy," to be discussed more extensively below.

It might be possible to make something similar to the integrative diagram on the level of dynamics rather than structures, complementing the structural integrative diagram given here; but this would seem significantly more challenging, because we lack a standard set of tools for depicting system dynamics. Most cognitive theorists and AGI architects describe their structural ideas using boxes-and-lines diagrams of some sort, but there is no standard method for depicting complex system dynamics. So to make a dynamical analogue to the integrative diagram, via a similar integrative methodology, one would first need to create appropriate diagrammatic formalizations of the dynamics of the various cognitive theories being integrated – a fascinating but onerous task.

When we first set out to make an integrated cognitive architecture diagram, via combining the complementary insights of various cognitive science and AGI theorists, we weren't sure how well it would work. But now we feel the experiment was generally a success – the resultant integrated architecture seems sensible and coherent, and reasonably complete. It doesn't come close to telling you everything you need to know to understand or implement a human-like mind – but it tells you the various processes and structures you need to deal with, and which of their interrelations are most critical. And, perhaps just as importantly, it gives a concrete way of understanding the insights of a specific but fairly diverse set of cognitive science and AGI theorists as complementary rather than contradictory.

1.5. Cognitive Synergy

The architecture of the mind, ultimately, has no meaning without an associated *dynamics*. Architecture emerges from dynamics, and channels dynamics. The cognitive dynamics of AGI systems is a large topic which we won't attempt to thoroughly pursue here, but we will mention one dynamical principle that we feel is essential for properly interpreting the integrative diagram: cognitive synergy.

Cognitive synergy has been proposed as a "general principle of feasible general intelligence".[?] It is both a conceptual hypothesis about the structure of generally intelligent systems in certain classes of environments, and a design principle that one may used to guide the architecting of AGI systems.

First we review how cognitive synergy has been previously developed in the context of "multi-memory systems" – i.e., in the context of the diagram given above for long-term memory and deliberative processing, Figure 1.4. In this context, the cognitive synergy hypothesis states that human-like, human-level intelligent systems possess a combination of environment, embodiment and motivational system that makes it important for them to possess memories that divide into partially but not wholly distinct components corresponding to the categories such as:

- Declarative memory
- Procedural memory (memory about how to do certain things)
- Sensory and episodic memory
- Attentional memory (knowledge about what to pay attention to in what contexts

- Intentional memory (knowledge about the system's own goals and subgoals)

The essential idea of cognitive synergy, in the context of multi-memory systems possessing the above memory types, may be expressed in terms of the following points:

(1) Intelligence, relative to a certain set of environments, may be understood as the capability to achieve complex goals in these environments.
(2) With respect to certain classes of goals and environments, an intelligent system requires a "multi-memory" architecture, meaning the possession of a number of specialized yet interconnected knowledge types, including: declarative, procedural, attentional, sensory, episodic and intentional (goal-related). These knowledge types may be viewed as different sorts of pattern that a system recognizes in itself and its environment.
(3) Such a system must possess knowledge creation (i.e. pattern recognition / formation) mechanisms corresponding to each of these memory types. These mechanisms are also called "cognitive processes."
(4) Each of these cognitive processes, to be effective, must have the capability to recognize when it lacks the information to perform effectively on its own; and in this case, to dynamically and interactively draw information from knowledge creation mechanisms dealing with other types of knowledge
(5) This cross-mechanism interaction must have the result of enabling the knowledge creation mechanisms to perform much more effectively in combination than they would if operated non-interactively. This is "cognitive synergy."

Interactions as mentioned in Points 4 and 5 in the above list are the real conceptual meat of the cognitive synergy idea. One way to express the key idea here is that most AI algorithms suffer from combinatorial explosions: the number of possible elements to be combined in a synthesis or analysis is just too great, and the algorithms are unable to filter through all the possibilities, given the lack of intrinsic constraint that comes along with a "general intelligence" context (as opposed to a narrow-AI problem like chess-playing, where the context is constrained and hence restricts the scope of possible combinations that needs to be considered). In an AGI architecture based on cognitive synergy, the different learning mechanisms must be designed specifically to interact in such a way as to palliate each others' combinatorial explosions - so that, for instance, each learning mechanism dealing with a certain sort of knowledge, must synergize with learning mechanisms dealing with the other sorts of knowledge, in a way that decreases the severity of combinatorial explosion.

One prerequisite for cognitive synergy to work is that each learning mechanism must recognize when it is "stuck," meaning it's in a situation where it has inadequate information to make a confident judgment about what steps to take next. Then, when it does recognize that it's stuck, it may request help from other, complementary cognitive mechanisms.

The key point we wish to make here regarding cognitive synergy is that this same principle, previously articulated mainly in the context of deliberative processes acting on

long-term memory, seems intuitively to hold on the level of the integrative diagram. Most likely, cognitive synergy holds not only between the learning algorithms associated with different memory systems, but also between the dynamical processes associated with different large-scale components, such as are depicted in the different sub-diagrams of the integrative diagram depicted above. If this is so, then all the subdiagrams depend intimately on each other in a dynamic sense, meaning that the processes within each of them must be attuned to the processes within each of the others, in order for the whole system to operate effectively. We do not have a proof of this hypothesis at present, so we present it as our intuitive judgment based on informal integration of a wide variety of evidence from cognitive science, neuroscience and artificial intelligence.

Of course, some pieces of the integrative diagram are bound to be more critical than others. Removing the language box might result in an AGI system with the level of intelligence of a great ape rather than a human, whereas removing significant portion of the perception box might have direr consequences. Removing episodic memory might yield behavior similar to certain humans with brain lesions, whereas removing procedural memory would more likely yield an agent with a basic inability to act in the world. But the point of cognitive synergy is not just that all the boxes in the integrative diagram are needed for human-level intelligence, but rather that the dynamics inside all the boxes need to interact closely in order to achieve human-level intelligence. For instance, it's not just removing the perception box that would harm the system's intelligence – forcing the perception box to operate dynamically in isolation from the processes inside the other boxes would have a similar effect.

1.6. Why Is It So Hard to Measure Partial Progress Toward Human-Level AGI?

Why it is so hard to measure partial progress toward human-level AGI? The reason is not so hard to understand, if one thinking in terms of cognitive synergy and system integration.

Supposing the integrative diagram is accurate – then why can't we get, say, 75% of the way to human level intelligence by implementing 75% of the boxes in the integrative diagram? The reason this doesn't work, we suggest, is that cognitive synergy possesses a frustrating but important property called "trickiness."

Trickiness has implications specifically for the evaluation of *partial* progress toward human-level AGI. It's not entirely straightforward to create tests to measure the *final achievement* of human-level AGI, but there are some fairly obvious candidates for evaluation methods. There's the Turing Test (fooling judges into believing you're human, in a text chat) the video Turing Test, the Robot College Student test (passing university, via being judged exactly the same way a human student would), etc. There's certainly no agreement on which is the most meaningful such goal to strive for, but there's broad agreement that a number of goals of this nature basically make sense.

On the other hand, it's much less clear how one should measure whether one is, say, 50 percent of the way to human-level AGI? Or, say, 75 or 25 percent?

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.

Some examples of practical tests of this nature would be

- The Wozniak "coffee test": 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)
- Graduating (virtual-world or robotic) preschool
- 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

One interesting point about tests like this is that each of them seems to *some* AGI researchers to encapsulate the crux of the AGI problem, and 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 people try really hard over a long period of time.

A question raised by these 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. 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?

We suggest that a partial answer is provided by the "trickiness" of cognitive synergy. Recall that, in its simplest form, the cognitive synergy hypothesis states that human-level AGI intrinsically depends on the synergetic interaction of multiple components. 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 almost surely 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. Or the point could be made without any decomposition into a finite set of components, using continuous probability distributions. To mathematically formalize the cognitive synergy hypothesis in its fully generality would become quite 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 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 supposition is a good bit stronger than mere cognitive synergy. For lack of a better name, we'll call it *tricky cognitive synergy*. The tricky cognitive synergy hypothesis would be true 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 AGI 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. In a CogPrime context, these possibilities ring true, in the sense that tailoring an AI process for tight integration with other AI processes within CogPrime, tends to require more work than preparing a conceptually similar AI process for use on its own or in a more task-specific narrow AI system.

It seems fairly obvious that, if tricky cognitive synergy really 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.

1.7. Conclusion

We have presented an integrative diagram summarizing and merging multiple researchers' views regarding the architecture of human-level general intelligence. We believe the results of our work demonstrate a strong degree of overlap and synergy between different

contemporary perspectives on AGI, and illustrate that a substantial plurality of the AGI field is moving toward consensus on the basic architecture of human-like general intelligence. Also, we suggest the integrative diagram may be useful from a purely cognitive science view, as a coherent high-level picture of all the parts of the human mind and how they work together.

We have also presented an argument that, to achieve anything remotely similar to human-level general intelligence, it will be necessary to implement *all* of the integrative diagram, not just isolated bits and pieces. We believe the arguments given here regarding trickiness provide a plausible explanation for the empirical observation that positing tests for intermediate progress toward human-level AGI is a very difficult prospect. If the theoretical notions sketched here are correct, then this difficulty is not due to incompetence or lack of imagination on the part of the AGI community, nor due to the primitive state of the AGI field, but is rather intrinsic to the subject matter. And in that case, the practical implication for AGI development is, very simply, that one shouldn't worry a lot about producing intermediary results that are compelling to skeptical observers. Just at 2/3 of a human brain may not be much use, similarly, 2/3 of an AGI system may not be much use. Lack of impressive intermediary results may not imply one is on a wrong development path; and comparison with narrow AI systems on specific tasks may be badly misleading as a gauge of incremental progress toward human-level AGI.

Thus our overall conclusion is both optimistic and pessimistic. If one implements a system instantiating the integrative diagram, and fills in each box with processes that cooperate synergetically with the processes in the other boxes to minimize combinatorial explosion – then one will get a human-level general intelligence. On the other hand, due to the trickiness of cognitive synergy, such a system may not display dramatic general intelligence until it is just about finished!

References

1. M. Minsky, *The Emotion Machine*. 2007.
2. A. Sloman. Varieties of affect and the cogaff architecture schema. In *Proceedings of the Symposium on Emotion, Cognition, and Affective Computing*, AISB-01, (2001).
3. B. Baars and S. Franklin, Consciousness is computational: The lida model of global workspace theory, *International Journal of Machine Consciousness.* (2009).
4. J. Bach, *Principles of Synthetic Intelligence*. (Oxford University Press, 2009).
5. D. Drner, *Die Mechanik des Seelenwagens. Eine neuronale Theorie der Handlungsregulation*. (Verlag Hans Huber, 2002). ISBN 345683814X.
6. J. S. Albus and A. M. Meystel, *Engineering of Mind: An Introduction to the Science of Intelligent Systems*. (Wiley and Sons, 2001).
7. I. Arel, D. Rose, and R. Coop, Destin: A scalable deep learning architecture with application to high-dimensional robust pattern recognition., *Proc. AAAI Workshop on Biologically Inspired Cognitive Architectures*. (2009).
8. J. Hawkins and S. Blakeslee, *On Intelligence*. (Brown Walker, 2006).
9. B. Goertzel. Opencog prime: A cognitive synergy based architecture for embodied artificial general intelligence. In *ICCI 2009, Hong Kong*, (2009).

10. G. Li, Z. Lou, L. Wang, X. Li, and W. J. Freeman, Application of chaotic neural model based on olfactory system on pattern recognition, *ICNC*. **1**, 378–381, (2005).
11. D. Jurafsky and J. Martin, *Speech and Language Processing*. (Pearson Prentice Hall, 2009).
12. B. e. a. Goertzel. A general intelligence oriented architecture for embodied natural language processing. In *Proc. of the Third Conf. on Artificial General Intelligence (AGI-10)*. Atlantis Press, (2010).
13. S. Franklin and B. Baars, Possible neural correlates of cognitive processes and modules from the lida model of cognition, *Cognitive Computing Research Group, University of Memphis*. (2008). `http://ccrg.cs.memphis.edu/tutorial/correlates.html`.
14. B. Goertzel. Cognitive synergy: A universal principle of feasible general intelligence? In *Proceedings of ICCI-09, Hong Kong*, (2009).