A Comparison of the Novamente AI Design with the Human Mind/Brain

Ben Goertzel
Novamente LLC
February, 2005

No one has tried to make a thinking machine….
The bottom line is that we really haven't progressed too far toward a truly intelligent machine.
We have collections of dumb specialists in small domains; the true majesty of general intelligence still awaits our attack.
…
We have got to get back to the deepest questions of AI and g e n e r a l intelligence and quit wasting time on little projects that don't contribute to the main goal."

-- Marvin Minsky

(as interviewed in *Hal's Legacy*, Edited by David Stork, 2000)

1 Introduction

The Novamente AI Engine is a novel software architecture that, unlike most contemporary AI projects, is specifically oriented towards artificial *general* intelligence (AGI), rather than being restricted by design to one particular domain, or narrow range of cognitive functions.

Novamente integrates aspects of prior AI projects and approaches, including symbolic, neural-network, evolutionary programming and reinforcement learning. However, its overall architecture is unique, drawing on system-theoretic ideas regarding complex mental dynamics and associated emergent patterns. Thus Novamente addresses the problem of "creating a whole mind" in a novel way through this integrative mechanism.

Novamente can be integrated with conventional software applications, enhancing their intelligence. It can also support radical new forms of human-computer interaction which includes a novel interface for mixed human/formal language conversation between humans and Novamente systems.

The overall mathematical and conceptual design of the Novamente AGI system is described in a series of manuscripts being prepared for publication in late 2005 or early 2006 (Goertzel and Pennachin, in prep.; Goertzel, Ikle' and Goertzel, in prep.; Goertzel, in prep.). The existing codebase implements roughly 60% of the design, and is being applied in bioinformatics and other domains.[1]

This article gives a high-level overview of the Novamente AI design, with a specific focus on comparing Novamente to the human brain/mind. Although Novamente's design was inspired in large part by cognitive science, it is not intended as a simulation of human intelligence; so this comparison is intended to reveal numerous differences as well as similarities. However, human intelligence is the best example of general intelligence we currently have, and for this reason the comparison seems very much worth drawing.

2 The Value of the Cognitive Sciences for AGI

The traditional disciplines of psychology and brain science, prior to the last few decades, offered extremely little to the would-be AGI designer. Since the emergence of cognitive science and cognitive neuroscience, things have gotten a little better. The state of knowledge in these disciplines is not yet sufficient to give detailed prescriptions for the construction of AGI systems. But these bodies of knowledge can provide substantial *inspiration* for AGI design.

On the one hand, the cognitive sciences provide very clear advice regarding what the overall "conceptual architecture" of an AGI system should be like, if that AGI system is going to cognize in a manner even vaguely resembling that of human beings. We know what the major regions of the brain do, and we also have a decent working decomposition of human cognition into a list of interacting yet significantly distinct faculties. This high-level architecture can be emulated in AGI systems.

On the other hand, the cognitive sciences provide a variety of suggestions regarding specific low-level mechanisms for carrying out intelligent processing, such as, perception, learning and memory. However, the low-level messages from the cognitive sciences are more controversial than the high-level ones for two reasons. First, there is less agreement on them among contemporary experts. And second, it's not always clear that emulating human psychological or neural behavior is a practical approach to implementing intelligence on radically un-brain-like hardware.

Cognitive theorists recognize that there is more to the human mind/brain than its high-level architecture and low-level mechanisms. However, the cognitive sciences to date have had relatively little to day about the crucial "intermediate level" of intelligence. This is the main reason that the cognitive sciences don't yet provide really powerful prescriptive guidance to AGI designers. The cognitive sciences tell us what major parts a mind/brain should have, and they describe some low-level mechanisms that can help

---

[1] The reason the development system has been so slow has been quite simple: lack of resources. There are no staff currently devoted to Novamente as an AGI system; rather, AGI development has been done by scientists and engineers working on Novamente-based narrow-AI applications, working in their "spare time."

these parts to carry out their functions, but they say precious little about how the different parts all work together, and how the low-level mechanisms coordinate to give rise to higher-level dynamics.

As an example of the extreme paucity of intermediate-level explanations, consider the rather critical notion of the "self." Thomas Metzinger (2004) has recently given a masterful treatment of the philosophy and neuropsychology of self, and has argued convincingly that a vast majority of the learning, thinking, perceiving and remembering that we humans do occurs in the context of the "phenomenal selves" that we construct in order to model ourselves and our relationships with the world. But as Metzinger points out, while contemporary cognitive neuroscience tells us a lot about various dysfunctions of self-construction and self-awareness, it doesn't say hardly anything about how selves are built by mind/brains. We know the self-building process involves the coordinated activity of a number of different brain regions, and we can list these; and we know some basic neural mechanisms that assist with the process, and can be diverted from their normal behavior via disturbances in the levels of various chemicals in the brain. But what does this "coordinated activity" consist of? How are chemical and neuron-level processes orchestrated across various brain regions to allow the human brain to model the mind that emerges from it in an approximately yet contextually-very-useful way? On issues like this, the cognitive sciences are basically silent.

In the world of cognitive neuroscience, controversy continues regarding the "binding problem" (Singer, 2001), the question of how all the neuro-sensory traces corresponding to different parts of an observed image are bound together in the brain to form a common, holistically perceived percept. Of course this is an important thing to be thinking about – but, when this kind of basic issue is still under dispute, it should be clear that, barring a sudden breakthrough, we're not all that close to understanding how cognitive traces reflecting memories and inferences about one's own self are bound into a unified self-system.

Next, as a much simpler example of the paucity of intermediate-level explanations, consider the issue of the relevance of temporal pattern-recognition to visual object-recognition. Jeff Hawkins (2004) has suggested that object recognition is based on hierarchical recognition of time series of visual sensations. The more traditional view argues that, while temporal analysis isn't totally irrelevant, it's not critical to the ordinary object recognition process? Who's right? No one knows. Most likely one could build computer vision systems based on either approach. The known architecture of the visual cortex supports either approach, as do the basic neurocognitive dynamics of Hebbian learning and neural-net activation spreading. This is a fairly simple, fairly low-level issue of how basic learning/memory mechanisms combine to yield the function known to be associated with a particular brain region of relatively well-known architecture. This is an easier issue than Metzinger's question about the neural basis of the construction of the phenomenal self – but again it illustrates the type of explanation that the cognitive sciences currently provide only very infrequently.

Finally, consider an issue relating to logical reasoning. At their best, humans are capable of carrying out highly complex trains of logical reasoning – including e.g. mathematical proofs, carefully orchestrated multi-part criminal bank fraud operations, and the construction of the concept of "logical reasoning" itself. Yet, even intelligent

humans routinely perform poorly on simple reasoning puzzles such as the Wason card task (Wason, 1966). Why? It's not that we're stupid, it's that our capability for logical reasoning is integrated into our overall mind-structure in a particular way that only allows it to display its full potential under certain conditions. Psychological experiments have shown that the propensity for accurate reasoning depends highly on various phenomena such as the familiarity of the domain being reasoned about. But yet, mathematically trained individuals will almost never make errors like the one demonstrated in the typical response to the Wason card task. Again: why? Are mathematically trained individuals using a different reasoning algorithm than ordinary people, or are they just "tuning" the universal reasoning algorithm in a different way, or connecting the universal reasoning subsystem to other brain subsystems in a different way, etc? Experimental psychology and neuroscience will approach this issue slowly and indirectly over the next decades, but right now an answer is nowhere near.

2.1 What's the AGI Designer To Do?

Given the current state of the cognitive sciences, the present-day AGI designer has several recourses.

Firstly, he can simply wait until the cognitive sciences advance further, and give more thorough prescriptions for AGI design.

Secondly, he can ignore the cognitive sciences and attempt to design an AGI on other grounds – e.g. based on the mathematics of reasoning, or based on general considerations regarding the dynamics of complex self-organizing systems. Of course it's worth reflecting that many of these "other grounds" –such as mathematical logic -- were originally conceived as qualitative models of human thought. But still, in spite of this historical fact and the strong intuitive feelings associated with it, the empirical cognitive sciences have not yet substantiated any deep connections between mathematical logic and human cognition.

Or, thirdly, he can seek to create an AGI design that is consistent with the information provided by the cognitive sciences, but also introduces additional ideas filling in the gaps they leave. This approach has a great deal of flexibility, of course. This is the approach we've taken in designing the Novamente AI system.

The overall goal of the Novamente AI project is *not* to create a human-like digital mind. While the human brain/mind is impressive, it has serious flaws with which we're all quite familiar (Pietelli-Palmarini, 1996), and we don't feel that replicating these flaws in software is a particularly desirable goal. Also, it's clear that the nature of human intelligence has been strongly influenced by the particularities of human embodiment – and similarly, we'd expect a digital mind to be strongly influenced by the particularities of its own embodiment, which (barring huge advances in android technology) will be quite different from that of humans. These caveats aside, however, we have a great deal of respect for the fact that the human brain/mind actually exists and functions, so we feel it would be foolish not to learn from it all that we can.

The high-level architecture of the Novamente system is closely inspired by well-known facts about the high-level architecture of human cognition. The particular learning, reasoning and perceptual mechanisms within Novamente are inspired by human

psychology and neuroscience, but more loosely. At this level, rather than seeking to emulate the details of how humans do things, we have sought to emulate the *spirit* of human processing. This choice was made largely with computational efficiency in mind. The brain contains a massive number of fairly slow and noisy processing units, and has an architecture in which memory and processing are largely overlapping concepts; on the other hand, modern computer networks have a small number of very fast and accurate processors, and an architecture in which memory and processing are distinct. These differences mean that the mechanisms that the brain has evolved, to make effective use of the physical wetware allocated to it, are not very well suited to efficient digital computer implementation. From an AGI point of view, until/unless radically more brainlike computer hardware comes along, the most sensible AGI cognitive-microdesign strategy seems to be to try to understand the essence of each human cognitive function, then figure out a way to implement this essence in a digital-computer-friendly way, without worrying about exactly how this function is implemented in the brain via neurons, neurotransmitters, extracellular charge diffusion and the like.

To bridge the gap between the high-level architecture and the low-level cognitive mechanisms, Novamente makes use of a novel theoretical approach called SMEPH, for "Self-Modifying, Evolving Probabilistic Hypergraphs". The SMEPH approach provides a mathematical and conceptual framework in which answers to questions such as "How does the phenomenal self emerge?" or "Does object recognition use hierarchical pattern recognition on temporal or static information?" can be crisply formulated, analyzed, and discussed. It doesn't, in itself, resolve all these questions – it just gives an approach to exploring solutions. The Novamente design applies the SMEPH approach in particular ways, which are intended to provide adequate though almost surely not optimal solutions to the main problems of intermediate-level AGI mind-design.

2.2 Competing Approaches to AGI and their Grounding in the Cognitive Sciences

Before turning to Novamente, it's worth briefly discussing how competing approaches to AGI relate to contemporary knowledge from the cognitive sciences.

The majority of software systems with explicit AGI ambitions are founded more directly on cognitive science than cognitive neuroscience. The reason for this is the obvious one noted above: neuroscience doesn't really give enough guidance. On the other hand, at least cognitive psychology says *something* about every major aspect of intelligence, even though its pronouncements aren't always convincing.

There are a number of narrow-AI systems based on qualitative models of neurodynamics – these are the well-known "neural network" algorithms, which have proved very useful in a variety of application domains, and have led to a lot of interesting mathematical theory (see e.g. Amit, 1989 for classical results). However, most neural net systems are basically using crude models of the microstructure and microdynamics of the brain to carry out particular learning or memory functions. They're not trying to emulate the overall structure of the brain as it gives rise to unified general intelligence.

One well-known exception is Stephen Grossberg's work (Grossberg, 1987 is the classic reference), which involves a collection of reasonably accurate neural net models of particular brain regions. Another is Peter Voss's (2005) A2I2 architecture, which is loosely based on the "neural gas" approach to neural net modeling, but uses a number of innovative structures and algorithms in a quest to make a specially-structured neural network learn from experience in a simple simulated environment, in a manner similar to an intelligent nonhuman mammal or a young child. Another exception is Hugo de Garis's CAM-Brain project (De Garis and Korkin, 2002), which is based on a combination of brain-based ideas and more computer-science-based AI concepts: in the CAM-Brain based "RoboKoneko" design, small neural networks are trained using a genetic algorithm and then arranged in a network-of-networks to carry out robot perception and control functions.

The best-funded loosely-AGI-oriented effort currently underway is almost surely the Cyc project (Lenat and Guha, 1999). Cyc is based on a particular theory of human intelligence which holds that *declarative knowledge* is the most important aspect of mind. Doug Lenat began Cyc with clear AGI ambitions, but in the decades since the project's inception, the Cyc team has in fact spent nearly all of their effort building a knowledge base and supplying it with logical inference tools, rather than creating a coherent, integrated digital intelligence. Currently there is a "Cognitive Cyc" project underway within Cycorp, whose aim is to move Cyc development in more of a holistic-AGI direction based on their existing knowledge base and reasoning engine.[2]

SOAR (Laird et al, 1987) and ACT-R (Anderson et al, 1997) are two cognitive-psychology-inspired AI systems that aim to model integrated intelligence. SOAR has been used to simulate the behavior of human pilots, with moderate success (Jones et al, 1993). These systems focus mainly on modeling human memory and logical problem-solving, but even in this domain their scope is quite limited.

There are also some innovative AGI-oriented systems founded on advanced theories of human logical reasoning. Pei Wang's (1995) NARS inference engine is based on a novel form of uncertain term logic. Stuart Shapiro's Sneps system (2000) is based on a variety of paraconsistent logic, and has been used (Santore et al, 2003) to control simple behaviors of an agent in a simulated world.

Jason Hutchens' HAL chat system (see www.a-i.com) seeks to emulate human development psychology – with a focus on language acquisition -- using an underlying statistical learning methodology. There are also a few robotics projects, e.g. Rodney Brooks' Cog work (Adams et al, 2000), that aim at AGI in the long term. While not based on detailed models of the brain, Brooks' work seeks to qualitatively emulate the way perception, action and decision occur in biological systems.

In all, we see that existing AGI projects have all made use of inspiration from the cognitive sciences in various ways. None has made a serious attempt to simulate brain function because not enough is known about the latter. Some, notably SOAR and ACT-R, have made serious attempts to emulate human psychological function, but even so they have left out a lot of extremely important components.

The Novamente project is unique among efforts at AGI in that it possesses a concrete and detailed mathematical, conceptual and software design, that provides a

---

[2] Stephen Reed, Cycorp; personal communication

unified treatment of all major aspects of intelligence as detailed in cognitive science and computer science.

3 Conceptual Underpinnings of Novamente

The primary motivation behind the Novamente AI Engine is to build a software system that can achieve complex goals in complex environments, a synopsis of the definition of intelligence given in (Goertzel 1993). The emphasis is on the plurality of *goals* and *environments*. A chess-playing program is not a general intelligence, nor is a data mining engine, nor is a program that can cleverly manipulate a researcher-constructed microworld. A general intelligence must be able to carry out a variety of different tasks in a variety of different contexts, generalizing knowledge between contexts and building up a context and task independent pragmatic understanding of itself and the world.

The Novamente design is founded on a philosophy of mind called the "psynet model," and on the SMEPH mathematical framework for modeling intelligent systems, which ties in naturally with the psynet model. But although these conceptual and formal tools were developed largely in the context of AI design, in fact they are general in nature, and have insight to shed into the nature of human intelligence as well. We reviewe them here both in preparation for the discussion of Novamente, and as a foundation for the principled comparison of Novamente with human cognition.

One aspect of Novamente's conceptual foundations that we will emphasize here is the notion of experiential interactive learning. The concepts here is that intelligence most naturally emerges through situated experience. Abstract thoughts and representations are facilitated through the recognition and manipulation of patterns in environments with which a system has sensorimotor interaction; see for example (Boroditsky and Ramscar 2002) for some elaborations on this notion. This philosophy implies that, from a Novamente perspective, a project such as embodiment in a robot or a simulated agent is more fundamentally interesting than for instance work on disembodied natural language processing.

3.1 The Psynet Model of Mind

The abstract principles underlying the Novamente architecture are coherently unified in a philosophy of cognition called the *psynet model* (Goertzel, 1993, 1993a, 1994, 1997, 2002), which provides a moderately detailed theory of the emergent structures and dynamics in intelligent systems. In the model, mental functions such as perception, action, reasoning and procedure learning are described in terms of interactions between agents. Any mind, at a given interval of time, is assumed to have a particular goal system, which may be expressed explicitly and/or implicitly. Thus, the dynamics of a cognitive system are understood to be governed by two main forces: self-organization and goal-oriented behavior.

More specifically, several primary dynamical principles are posited, including:

- **Association.** Patterns, when given attention, spread some of this attention to other patterns that they have previously been associated with in some way. Furthermore, there is Peirce's "law of mind" (Peirce, 1892), which could be paraphrased in modern terms as stating that the mind is an associative memory network, whose dynamics dictate that every idea in the memory is an active agent, continually acting on those ideas with which the memory associates it.
- **Differential attention allocation.** Patterns that have been valuable for goal-achievement are given more attention, and are encouraged to participate in giving rise to new patterns.
- **Pattern creation.** Patterns that have been valuable for goal-achievement are mutated and combined with each other to yield new patterns.
- **Credit Assignment.** Habitual patterns in the system that are found valuable for goal-achievement are explicitly reinforced and made more habitual.

Furthermore, the network of patterns in the system must give rise to the following large-scale emergent structures

- **Hierarchical network.** Patterns are habitually in relations of control over other patterns that represent more specialized aspects of themselves.
- **Heterarchical network.** The system retains a memory of which patterns have previously been associated with each other in any way.
- **Dual network.** Hierarchical and heterarchical structures are combined, with the dynamics of the two structures working together harmoniously.
- **Self structure.** A portion of the network of patterns forms into an approximate image of the overall network of patterns.

The key to the implementation of these general principles in a practical, mathematically sound AGI software design is the SMEPH framework for modeling intelligent systems in terms of self-modifying, evolving probabilistic hypergraphs, which will be described briefly below.

3.1.1 The Psynet Model and the Human Mind-Brain

While the psynet model is being considered here mainly in its role as a motivation for the Novamente AI design, as originally presented it was applied equally much to human psychology as to AI. Goertzel (1997) sketches compact arguments in favor of the psynet model as a conceptual model of human psychology and the structure of the human neocortex. Some of these ideas will be discussed in detail below when Novamente and the human brain/mind are contrasted based on modern cognitive-science theories and data. But the general nature of the correspondence is not hard to see.

Hierarchical structures pervade the brain – the cortex is hierarchical, visual perception has been shown to use this physical hierarchy to do hierarchical pattern recognition, and a number of theorists have proposed that this principle applies more

generally (see Hawkins, 2004). On the other hand the prevalence of association structures in the brain has been well-known since the time of Peirce and William James, and has been validated via numerous experiments on the structure of memory (Baddeley, 1999). The dual network as a general model of the brain's "concept/percept/action interconnection statistics" is not convincingly proven but is highly plausible. The way that self-structures may emerge from dual networks based on experiential learning has been discussed in depth in (Goertzel, 1997), with many connections drawn to the literature on personality psychology, e.g. Rowan's (1990) theory of subpersonalities.

The model of the human brain/mind as a system focused on pattern recognition and creation is also reasonably well-established, although it has not been systematically articulated as often one would like. Again, Peirce and James established this perspective long ago (though using the language of "habits" rather than "patterns"), and it has more recently reared its head in the form of algorithmic-information-theoretic models of the mind as "compact programs for computing data" (see Solomonoff, 1964, 1964a; or more recently, Baum 2004). The mathematical theory of pattern given in (Goertzel, 1997) shows how algorithmic information theoretic models are substantially the same as models based on concepts like pattern and habit.

Putting the mind-as-habit-system theme together with the dual-network structure – both general principles shown to be harmonious with contemporary cognitive science – one obtains a crude argument for the relevance of the psynet model to the human mind-brain. The next question becomes one of dynamics. What do all these patterns do, in order to help a mind achieve its goals and maintain its dual network structure? The psynet model posits a kind of evolutionary dynamic on the level of patterns: patterns found useful for system goals are reinforced, and combined with each other. This is a form of "evolutionary learning" (Holland, 1992), which Edelman (1987) and others have presented as a model of cognition. If we accept Edelman's views we then may portray the human mind-brain as an evolving dual network of patterns – i.e., a psynet.

This of course is a speculative, high-level, system-theoretic portrayal of the human brain/mind. However, we believe this is the level of abstraction one must begin with if one wishes to create an AGI system using inspiration from the cognitive sciences. The knowledge generated by the cognitive sciences so far is too primitive and too haphazard to be used as a blueprint for AI. One first of all needs a common conceptual framework for analyzing both human and AI cognitive structures and dynamics – and then within this framework, one can see how specific facts from cognitive science may be useful for guiding AGI development.

3.2 Experiential Interactive Learning

Based on the premise that a mind is the set of patterns in a brain, the psynet model describes a specific set of high-level structures and dynamics for mind-patterns, and proposes that these are essential to any sort of mind, human or digital. These are not structures that can be programmed into a system; rather they are structures that emerge through the situated evolution of a system – through experiential interactive learning.

Novamente's specific structures and dynamics are based on the more general ones posited by the psynet model.

The psynet model also contains a theory of the relation between mind, body and society that contrasts with the most common perspectives expressed in the AI literature. Namely, it maintains that s*oftware and mathematics alone, no matter how advanced, cannot create an AGI.* What software and mathematics *can* do, however, is to create an environment within which artificial general intelligence *emerges* through interaction with humans in the context of a rich stream of real-world data. That is: *Intelligence most naturally emerges through situated and social experience.*

It is clear that human intelligence does not emerge solely through human neural wetware. A human infant is not so intelligent, and an infant raised without proper socialization will never achieve full human intelligence (Douthwaite, 1997). Human brains learn to think through being taught, and through diverse social interactions. Our experience is that the situation is similar with AGI's. The basic AGI algorithms in Novamente are not quite adequate for practical general intelligence, because they give only the "raw materials" of thought. What is missing in Novamente "out of the box" are context-specific control mechanisms for the diverse cognitive mechanisms. The system has the capability to learn these, but just as critically, it has the capability to *learn how to learn* these, through social interaction.

A Novamente "out of the box" will be much "smarter" than narrow AI systems, but not as robustly intelligent as a Novamente that has refined its ability to learn context-specific control mechanisms through meaningful interactions with other minds. For instance, once it's been interacting in the world for a while, it will gain a sense of how to reason about conversations, how to reason about network intrusion data, how to reason about bio-warfare data – by learning context-dependent inference control schemata for each case, according to a schema learning process tuned through experiential interaction.

This leads us to the concepts of *autonomy*, *experiential interactive learning or EIL*, and *goal-oriented self-modification* – concepts that lie right at the heart of the notion of Artificial General Intelligence.

An integrative AI software system *may* be supplied with specific, purpose-oriented control processes and in this way used as a data mining and/or query processing engine. This is the approach taken, for example, in the current applications of the Novamente engine in the bioinformatics domain. But this kind of deployment of Novamente does not permit it to develop its maximum level of general intelligence.

For truly significant AGI to emerge, an emergent system must be supplied with general goals, and then allowed to learn its own control processes via execution of its procedural learning dynamics through interaction with a richly structured environment along with extensive meaningful interactions with other minds.

The Novamente system will gain its intelligence through processing relevant data, interacting with humans' in the context of this data, and providing humans with reports summarizing patterns it has observed. In this process, it will do more than increase its knowledge store, it will learn how to learn, and learn about itself. It will continually modify its control schemata based on what it's learned from its environment and the humans it interacts with. This process of "experiential interactive learning" has been one of the primary considerations in Novamente design and development.

While conversations about useful information will be an important source of EIL for Novamente, we suspect that additional tutoring on basic world-concepts like objects, motions, self and others will be valuable. For this purpose we have created a special simulated environment for the purpose of instructing Novamente: the AGI-SIM simulation world.

3.2.1 AGI-SIM

AGI-SIM is being developed as an open-source project with the intention of being useful for other AGI projects as well as Novamente. It is based on the open-source 3D simulation environment CrystalSpace[3]. Without going into details on AGI-SIM here, it is worth mentioning some of the basic principles that went into its design.

- The experience of an AGI, controlling an agent in a simulated world, should display the main qualitative properties of a human controlling their body in the physical world. For specific consideration are those qualitative properties which help the AGI to relate experiences in the simulated world to the many obvious and subtle real-world metaphors embedded in human language.

- The simulation world should support the integration of perception, action and cognition in a unified learning *loop*, which is crucial to the development of intelligence.

- The sim world should support the integration of information from a number of different senses, all reporting different aspects of a common world, which is valuable for the development of intelligence.

With these goals in mind, we have created AGI-SIM as a basic 3D simulation of the interior of a building, with simulations of sight, sound, smell and taste. An agent in AGI-SIM has a certain amount of energy, and can move around and pick up objects and build things. The initial version doesn't attempt to simulate realistic physics, but this may be integrated into a later version using the ODE open-source physics simulation package. While not an exact simulation of any physical robot, the agent Novamente controls in AGI-SIM is designed to bear enough resemblance to a simple physical robot that the same control routines should be portable to a physical robot – which is a step we look forward to taking, but feel is best postponed until after Novamente's learning in the simulation world is fairly far advanced.

3.3 Self-Modifying, Evolving Probabilistic Hypergraphs

Now how does one move from these conceptual generalities in the direction of a concrete design for an AGI system? There are many different AGI systems consistent with the general principles of the psynet model. As an intermediate level between the psynet model and specific AGI system designs, we have developed a mathematical

[3] **crystal**.sourceforge.net/

framework for modeling intelligent systems that we call the SMEPH (Self-Modifying, Evolving Probabilistic Hypergraph) approach. Novamente is based on SMEPH, but it is not the only way to make an AI system based on SMEPH; another example is the Webmind AI Engine (Goertzel et al, 2000) developed in the late 1990's. Furthermore, SMEPH is intended to be useful for analyzing intelligences that are not explicitly architected with SMEPH in mind. For example, the Hebbian Logic AI design outlined in (Goertzel, 2003) is based on neural net ideas rather than explicitly involving SMEPH data structures and dynamics, but its emergent structures and dynamics are naturally modeled in SMEPH terms. And many of the ideas in (Goertzel, 1993, 1993a, 1994, 1997) may be viewed as applications of SMEPH ideas to the analysis of human cognition (although the SMEPH framework had not been fully formalized and abstracted at that point).

SMEPH can be divided into two aspects: a way of representing knowledge (the Probabilistic Hypergraph part), and a way of representing dynamics (the Self-Modifying, Evolving part). In psynet model terms, probabilistic hypergraphs may be viewed as a particular way of representing patterns. Self-modification and evolution are key aspects of pattern dynamics, which are naturally representable in terms of the hypergraph knowledge representation. Here we will give a brief overview of SMEPH concepts and discuss their applicability to the human brain/mind and to Novamente.

3.3.1 Derived Hypergraphs

Now we will explain how, in the SMEPH framework, a complex system in an environment can be associated with a "derived hypergraph" capturing important aspects of the structure and dynamics of the system.

Before getting started, however, a brief note on terminology is necessary. In mathematics, one can choose to refer to a graph as consisting of *nodes and links*, or else as *vertices and edges*. The two pairs of terms mean the same thing. However, in talking about SMEPH and Novamente, we will use the terms in different ways. We will reserve "nodes and links" for talking about objects in Novamente's explicit knowledge representation (which is a hypergraph). On the other hand, we will use "vertices and edges" to talk about parts of an abstract SMEPH hypergraph. This becomes subtle because SMEPH may be used to model Novamente, and according to this modeling, sometimes Novamente nodes/links may correspond to SMEPH vertices/edges, but other times it will be fuzzy sets of Novamente nodes/links (called "maps") that correspond to SMEPH vertices/edges.

A hypergraph, in mathematics, typically refers to a graph in which edges can span more than two vertices (Berge, 1999). Hypergraphs as we consider them are even more general than that, as we consider edges that can point to edges as well as vertices. Also, we consider weighted hypergraphs, in which edges and vertices come along with sets of numbers with particular semantics. In particular, SMEPH specifies that each edge or vertex must be associated with a package of numbers called an AttentionValue, specifying how much attention has been paid to it (there may be several different AttentionValue numbers corresponding to different time scales. Novamente uses two, called (short-term) "importance" and "long-term importance"). Also, many edges and

vertices will be associated with packages of numbers called TruthValues, which indicate the probability with which the relationship denoted by the edge or vertex is true. The semantics of an edge or vertex's TruthValue depends on the particular type of edge or vertex, which brings us to a final feature of SMEPH's formal knowledge representation: edges or vertices may have types, drawn from a finite set. There is a minimal set of edge and vertex types associated with SMEPH; modeling particular intelligent systems explicitly based on the SMEPH approach may make it natural to introduce additional edge or vertex types on a case-by-case basis.

The basic vertex types contained in the SMEPH approach are: Concept and Schema. These are terms overloaded with meanings, and we use them in SMEPH as defined terms, without pretending that our usage accords fully with all natural language usages.

A Concept, in SMEPH, refers to the habitual pattern of activity observed in a system when some condition is met. The condition may refer to something in the world external to the system, or to something internal. For instance, the condition may be "observing a cat." In this case, the corresponding Concept vertex in the mind of Ben Goertzel is the pattern of activity observed in Ben Goertzel's brain when his eyes are open and he's looking in the direction of a cat. The notion of "pattern of activity" can be made rigorous using mathematical pattern theory (Goertzel, 1997).

Note that logical predicates, on the SMEPH level, appear as particular kinds of Concepts, where the condition involves a predicate and an argument. For instance, suppose one wants to know what happens inside Ben's mind when he eats cheese. Then there is a Concept corresponding to the condition of cheese-eating activity. But there may also be a Concept corresponding to eating activity in general. If the Concept for X-eating activity is generally easily computable from the Concepts for X and eating individually, then the eating Concept is effectively acting as a predicate.

A Schema, on the other hand, is like a Concept that's defined in a time-dependent way. One type of Schema refers to a habitual dynamical pattern of activity occurring before and/or during some condition is met. For instance, the condition might be saying the word "Hello." In that case the corresponding Schema vertex in the mind of Ben Goertzel is the pattern of activity that generally occurs before he says "Hello."

Another type of Schema refers to a habitual dynamical pattern of activity occurring after some condition X is met. For instance, in the case of the Schema for adding two numbers, the precondition X consists of the two numbers and the concept of addition. The Schema is then "what happens when the mind thinks of adding and thinks of two numbers."

Finally, there are Schema that refer to habitual dynamical activity patterns occurring after some condition X is met and before some condition Y is met. In this case the Schema is viewed as transforming X into Y. For instance, if X is the condition of meeting someone who is not a friend, and Y is the condition of being friends with that person, then the habitually intervening activities constitute the Schema for making friends.

SMEPH edge types fall into two categories: functional and logical. Functional edges connect Schema vertices to their input and outputs. The Execution edge denotes a relation between Schema, its input and its output, e.g.

```
Execution make_friends meets_Fred is_friend_of_Fred
```

The ExecutionOutput (ExOut) edge denotes the output of a Schema in an implicit way, e.g.

```
ExOut say_hello
```

refers to a particular act of saying hello, whereas

```
ExOut add_numbers {3, 4, "addition")
```

refers to the Concept corresponding to 7. Note that this latter example involves a set of three entities: sets are also part of the basic SMEPH knowledge representation. A set may be thought of as a hypergraph edge that points to all its members.

Logical edges refer to conditional probabilities: for instance, it may happen that whenever the Concept for "cat" is present in a system, the Concept for "animal" is as well. Then we would say

```
Subset cat animal
```

On the other hand, it may be that 50% of the time that "cat" is present in the system, "cute" is present as well: then we would say

```
Subset cat animal <.5>
```

where the <.5> denotes the probability, which is a component of the TruthValue associated with the edge. There is a collection of roughly a dozen different logical edge types in SMEPH, which are derived from the Probabilistic Term Logic framework (Goertzel, Ikle and Goertzel, in prep.). We will discuss some of these types in more depth in a later section, in the context of Novamente's closely related usage of PTL.

In this manner we may define a set of edges and vertices modeling the habitual activity patterns of a system when in different situations. This is called the "derived hypergraph" of the system. Note that this hypergraph can in principle be constructed no matter what happens inside the system: whether it's a human brain, a formal neural network, Cyc, Novamente, a quantum computer, etc. Of course, constructing the hypergraph in practice is quite a different story: for instance, we currently have no accurate way of measuring the habitual activity patterns inside the human brain. fMRI and PET technologies give only a crude view, though they are continually improving (Cabeza and Kingstone, 2001).

The psynet model comes in here and makes some definite hypotheses about the structure of derived hypergraphs. It suggests that derived hypergraphs should have a dual network structure, and that in highly intelligent systems they should have subgraphs that constitute models of the whole hypergraph (these are "self systems"). SMEPH does not add anything to the psynet model on a philosophical level, but it gives a concrete instantiation to the psynet model's general ideas.

3.3.2 Probabilistic and Evolutionary Dynamics

The logical edges in a SMEPH hypergraph are weighted with probabilities, as in the simple example given above. The functional edges may be probabilistically weighted as well, since some Schema may give certain results only some of the time. These probabilities are critical in terms of SMEPH's model of system dynamics; they underly one of SMEPH's three key principles of the dynamics of intelligence,

Principle of Implicit Probabilistic Inference: *In an intelligent system, the temporal evolution of the probabilities on the edges in the system's derived hypergraph should approximately obey the rules of probability theory.*

What "the rules of probability theory" means in this context is a complex issue and is addressed in (Goertzel et al, 2005a). The basic idea is that, even if a system's underlying dynamics has no explicit connection to probability theory, nevertheless it must behave roughly as if it does, if it is going to be intelligent. The "roughly" part is important here – it's well known that humans are not terribly accurate in explicitly carrying out formal probabilistic inferences. And yet, in practical contexts where they have experience, humans can make quite accurate judgments – which is all that's required by the above principle, since it's the contexts where experience has occurred that will make up a system's derived hypergraph.

The next key dynamical principle of SMEPH is evolutionary, and states

Principle of Implicit Evolution: *In an intelligent system, new Schema and Concepts will continually be created, and the Schema and Concepts that are more useful for achieving system goals (as demonstrated via probabilistic implication of goal achievement) will tend to survive longer.*

Note that this principle can be fulfilled in many different ways. The important thing is that system goals are allowed to serve as a selective force.

The final SMEPH dynamical principle pertains to a shorter time-scale than evolution, and states

Principle of Attention Allocation: *In an intelligent system, Schema and Concepts that are more useful for attaining short-term goals will tend to consume more of the system's energy*.

Derived hypergraphs may be constructed corresponding to any complex system which demonstrates a variety of internal dynamical patterns depending on its situation. However, if a system is not intelligent, the evolution of its derived hypergraph can't be expected to follow the above principles.

These principles follow from the psynet model of mind, but they are more precise than the psynet model can be, because they assume a particular formalism for representing the contents of a mind (SMEPH hypergraphs). Of course, no particular

mind will be completely described by this sort of hypergraph model; the idea is that this level of approximate description is good enough for many purposes.

The relationship between the human brain/mind and Novamente may be explored in a SMEPH context, by considering that both Novamente and the human mind can be modeled as SMEPH hypergraphs that obey the principles of implicit probabilistic inference and evolution. Below we will use this approach to help organize our discussion of various concrete results from the cognitive sciences and their relevance for Novamente and AGI in general.

3.3.3 From SMEPH to Novamente

While SMEPH is a general approach to modeling any intelligent system, it is also possible to create intelligent systems bearing a special relationship to SMEPH. Novamente falls into this category, as did Webmind. This special relationship makes it particularly easy to analyze these AI systems in SMEPH terms, but it also gives rise to potential confusions.

Novamente represents knowledge internally using a hypergraph data structure that involves nodes and links similar to SMEPH's edges and vertices. However, Novamente's vocabulary of node and link types is richer than SMEPH's, and the semantics of its nodes and links are different than that of SMEPH's edges and vertices. For instance, Novamente has node types called ConceptNode and SchemaNode, but also others like PredicateNode and various types of PerceptNodes. A Novamente ConceptNode will not generally represent a SMEPH Concept edge, because it's rare that Novamente's response to a situation will consist solely of activating a single ConceptNode. Rather, the Concept edges in the derived hypergraph of a Novamente system will generally correspond to fuzzy sets of Novamente nodes and links.

The term "map" is used in Novamente to refer to a fuzzy set of nodes and links that corresponds to a SMEPH concept or schema; and there is a typology of Novamente maps, to be briefly discussed below. Often it happens that a particular Novamente node will serve as the "center" of a map, so that e.g. the Concept edge denoting "cat" will consist of a number of nodes and links roughly centered around a ConceptNode that is linked to the WordNode "cat." But this is not guaranteed – some Novamente maps are more diffuse than this with no particular center.

Somewhat similarly, the key SMEPH dynamics are represented explicitly in Novamente: probabilistic reasoning is carried out via explicit application of PTL on the Novamente hypergraph, evolutionary learning is carried out via application of the BOA optimization algorithm, and attention allocation is carried out via a combination of inference and evolutionary pattern mining. But the SMEPH dynamics also occur implicitly in Novamente: emergent maps are reasoned on probabilistically as an indirect consequence of node-and-link level PTL activity; maps evolve as a consequence of the coordinated whole of Novamente dynamics; and attention shifts between maps according to complex emergent dynamics.

4 The Novamente AI Engine

The Novamente AI Engine is a unique, integrative AI architecture with general intelligence ambitions. It is a particular implementation of the SMEPH approach to intelligence-modeling, developed in such a way as to provide a framework for artificial general intelligence and also for the construction of commercial narrow AI applications. Implemented for efficiency and scalability on a distributed computing framework, Novamente uses two main AI tools – Probabilistic Term Logic (PTL) and the Bayesian Optimization Algorithm (BOA) -- to generate numerous cognitive processes operating on an evolving probabilistic hypergraph stored across multiple functionally specialized lobes.

The development of the Novamente system is not yet complete, and the functionality of Novamente as an AGI system remains unproven; however, the Novamente AI software framework has already proven itself in several practical applications. There is a pragmatic, application-driven path from the current state of Novamente toward the medium- and long-term AGI goals of the system. Table 1 gives a high-level portrayal of the project's development over time, both historically and in terms of projected future milestones (the achievement of which depends, of course, on a variety of factors including project funding).

4.1 Practical Applications of the Novamente AI Engine

The Novamente design is embodied in a C++ implementation, which is under active development. A number of performance issues, such as effectively swapping atoms between disk and memory, and distributed processing, have been dealt with via extensive optimization and testing. PTL and BOA have both been implemented and tested successfully; and much of the natural language framework has been completed.

In parallel with the development of Novamente toward the goal general intelligence, the system has been utilized more narrowly as an "AI toolkit" in the construction of practical commercial software applications, for example:

Bioinformatics. The Biomind Analyzer, developed by Biomind LLC together with Novamente LLC, is an enterprise system for intelligent analysis of microarray gene expression data. BOA is used to uncover interesting patterns in labeled datasets, and also to learn classification models. PTL is used for the integration of background information from a number of heterogeneous sources of biological knowledge, covering gene and protein function, research papers, gene sequence alignment, protein interactions, and pathways. This allows the Biomind Analyzer to augment the datasets it analyzes with background features corresponding to gene or protein categories, participation in pathways, etc. Inference is then used to create new relations between the genes and the functional categories provided by the background sources, effectively suggesting function

assignments to genes with unknown roles. For an example of work done using this process, see (Pennachin et al, 2005).

Milestone	Date	Description
M1	2001	Technical & Mathematical design + core implimentation
M2	2002	General quantitative data analysis *customization*
M2.1	2003	Gene expression microarray data analysis *customization*
M3	2004	Natural language processing *customization*
M3.1	2004	User interface for natural language entry *customization*
M4	2004	Combinator-BOA algorithm for procedure/predicate learning
M5	2004	Probabilistic inference module: First-order
M6	2004	Probabilistic inference module: Higher-order
M7	2004	Distributed architecture Version 1
M8	2005	Creation of AGI-SIM simulation world
M9	2005	Completion of probabilistic attention allocation
M10	2005	Agent control in AGI-SIM
M11	2006	Distributed architecture Version 2
M12	2006	Self-modification architecture Version 1

Novamente Development Milestones

Milestone	Date	Description
M1	2006	Experiential grounding of simple language understanding
M2	2006	Goal-directed navigation: the ability to find and retrieve objects
M3	2007	Collaborative and creative play in *Blocks World*
M4	2007	Ethical Behavior and Socialization

Novamente Teaching Milestones

Color Key: Completed | In Progress | Future

Table 1. Novamente Development Milestones. (Note: *customization* in this table refers to narrow-AI work done largely outside of the scope of AGI, although most customizations improve the system, to varying degrees, for AGI.)

Human Language Processing and Knowledge Management. The INLINK interactive knowledge entry framework (Goertzel et al, 2005), developed by Object Sciences Corp. together with Novamente LLC, utilizes Novamente's cognitive algorithms in combination with its NLP framework. Knowledge is entered via an interactive interface, which allows users to review and revise the system's understanding of that knowledge. A knowledge base is thus produced, which is augmented by reasoning, and may be queried in English or a special formal language. BOA Pattern Mining is used to spontaneously create queries that are judged interesting.

4.2 Knowledge Representation

The SMEPH framework does not tell you how to build a mind, only, in general terms, what a mind should be like. It would be possible to create many different AI designs based loosely on the psynet model and the SMEPH approach; one example of this is the Webmind AI Engine developed in the late 1990's (Goertzel et al. 2000, Goertzel 2002). Novamente, as a specific system inspired by these general ideas, owes many of its details to the limitations imposed by contemporary hardware performance and software design methodologies. Furthermore, Novamente is intended to utilize a minimal number of different knowledge representation structures and cognitive algorithms.

Knowledge representation in Novamente involves two levels, the explicit and the emergent: we will discuss both here, in sequence. Explicit knowledge representation utilizes a hypergraph formalism inspired by the SMEPH approach, which somewhat resembles classic semantic networks but has dynamic aspects that are more similar to neural networks. This enables a breadth of cognitive dynamics, but in a way that utilizes drastically less memory and processing than a more low-level, neural network style approach. The details of the representation have been designed for compatibility with the system's cognitive algorithms. Tables 3 and 4 review the key node and link types used in Novamente. Emergent knowledge representation, on the other hand, involves "maps" – fuzzy sets of nodes and links that respond to situations as habitual patterns, in the manner of SMEPH Concepts and Schema.

4.2.1 Explicit Knowledge Representation

Novamente's explicit knowledge hypergraph involves discrete units (called Atoms) of several types: Nodes, Links and Containers (the latter are ordered or unordered collections of atoms). Each Atom is associated with a TruthValue, indicating, roughly, the degree to which it correctly describes the world. Novamente has been designed with several different types of truth values in mind; the simplest of these consists of a pair of values denoting probability and weight of evidence. All Atoms also have an associated AttentionValue indicating how much computational effort should be expended on them. These contain two values, specifying short and long term importance levels.

Novamente node types include

- ConceptNodes, which derive their meaning via interrelationships with other nodes
- PerceptNodes nodes representing perceptual inputs into the system (e.g., pixels, points in time, etc.)
- TimeNodes representing moments and intervals of time
- PredicateNodes representing complex patterns
- SchemaNodes embodying procedures

SchemaNodes and PredicateNodes are nodes containing procedures that output Atoms and truth values, respectively. Procedures in Novamente are objects that produce an output, possibly based on a sequence of atoms as input. These objects contain structures called *generalized combinator trees* -- small computer programs written in *sloppy combinatory logic*, a language that we have developed specifically to meet the needs of tightly integrated inference and learning, utilizing ideas from combinatory logic as originally introduced in (Curry and Feys, 1958).

There are also special-purpose predicates that, instead of containing combinator trees, represent specific queries that report to the Novamente system some fact about its own state – these are called "feeling nodes". And finally, some predicates may also be designated as "goal nodes", in which case the system's GoalSatisfaction MindAgent allocates effort towards making them true.

Finally, Links are Atoms that represent relationships between other Atoms, such as fuzzy set membership, probabilistic logical relationships, implication, hypotheticality, and context. The complete list of (a few dozen) types and subtypes of links used, and the justifications for their inclusion, are omitted here for brevity. However, the most essential links are the Inheritance link (representing probabilistic logical implication), the Similarity link (a symmetric version of Inheritance), and the ExecutionOutputLink (representing the application of a function to an argument).

The network of Inheritance links forms an approximate hierarchical network, and the network of Similarity links forms a heterarchical network – the overlap between them forms a dual network structure, which emerges as a consequence of PTL reasoning operations that build new links from old. A dual network of maps forms implicitly from the dual network of nodes and links, thus forming the "emergent structure of mind" required according to the psynet model that is Novamente's conceptual foundation.

The semantics of logical links in Novamente are probabilistic, similar to in SMEPH. For example, we may write

```
InheritanceLink Iraq nation
```

meaning that there are ConceptNodes corresponding to the concepts "Iraq" and "nation," and there is an InheritanceLink pointing from one to the other (signifying that Iraq is indeed a nation). Or we may write

```
AssociativeLink Iraq terrorism <.7>
```

which just indicates a generic association between the two denoted ConceptNodes, with probabilistic strength .7. An associative relationship is useful for the spreading of

attention between related concepts, and also useful as a signpost telling the logical inference MindAgents where to look for possibly interesting relationships.

A more concrete relationship between Iraq and terrorism, such as "many terrorists live in Iraq," might be represented as

```
ImplicationLink lives_in_Iraq is_terrorist <.01>
```

where lives_in_Iraq and is_terrorist are PredicateNodes, and the former predicate obeys a relationship that would be written

```
EquivalenceLink (lives_in_Iraq (X)) (lives_in (Iraq , X) )
```

using variables, or

```
EquivalenceLink lives_in_Iraq  (lives_in (Iraq) )
```

using a variable-free, combinatory-logic-based internal representation (Novamente supports either representational style).

These have been examples of declarative knowledge; procedural knowledge, on the other hand, is represented via SchemaNodes and PredicateNodes, which embody snippets of code carrying out small procedures. The set of elementary schema/predicate functions is in effect an "internal Novamente programming language," which bears some resemblance to functional programming languages like pure LISP or Haskell. This language is represented externally as a language called Combo, which is difficult to write programs in due to the lack of local variables. There is also software allowing a more usable language called Sasha, a variant of the functional language Speagram (www.speagram.org), to compile into Combo.

4.2.2 Implicit Knowledge Representation

Much of the meaning of Novamente's cognitive algorithms lies in the implications they have for dynamics on the map level. Here the relation between Novamente Maps and the concepts of mathematical dynamical systems theory is highly pertinent.

Generally speaking there are two kinds of maps: map attractors, and map transients. Schema and predicate maps generally give rise to map transients, whereas concepts and percepts generally give rise to map attractors; but this is not a hard and fast rule. Other kinds of maps have more intrinsic dynamic variety, for instance there will be some feeling maps associated with transient dynamics, and others associated with attractor dynamics.

Many concept maps will correspond to fixed point map attractors – meaning that they are sets of Atoms which, once they become important, will tend to stay important for a while due to mutual reinforcement. However, some concept maps may correspond to more complex map dynamic patterns. And event maps may sometimes manifest a dynamical pattern imitating the event they represent. This kind of knowledge representation is well known in the attractor neural networks literature .

Schemata, on the other hand, generally correspond to transient maps. An individual SchemaNode does not necessarily represent an entire cognitive procedure of any significance – it may do so, especially in the case of a large encapsulated schema; but more often it will be part of a distributed schema. A distributed schema is a kind of mind map, and its map dynamic pattern is simply the system behavior that ensues when it is executes – behavior that may go beyond the actions explicitly embodied in the SchemaNodes contained in the distributed schema.

The maps in the system build up to form larger and more complex maps, ultimately yielding very large-scale emergent patterns, including patterns like the "dual network" (a combined hierarchical/heterarchical control structure) and the "self" (a fractal pattern in which a subnetwork of the hypergraph comes to resemble the hypergraph itself), which are posited in the psynet model of mind.

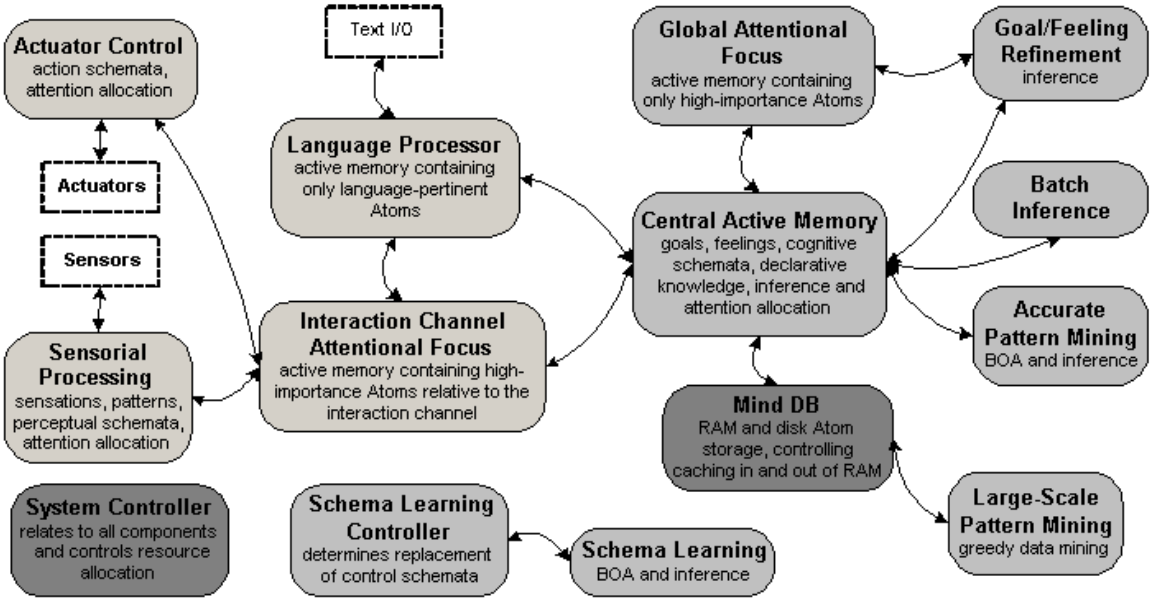

Figure 1. *High-level architecture of a complex Novamente instantiation. Each component is a Lobe, which contains multiple atom types and mind agents. Lobes may span multiple machines, and are controlled by schemata which may be adapted/replaced by new ones learned by Schema Learning, as decided by the Schema Learning Controller. The diagram shows a configuration with a single interaction channel, that contains sensors, actuators and linguistic input; real deployments may contain multiple channels, with different properties. (From Looks et al, 2004)*

4.3 Architecture and Dynamics

In Novamente we have reduced the set of fundamental cognitive algorithms to two: Probabilistic Term Logic (PTL) and the Bayesian Optimization Algorithm (BOA; see Pelikan, 2002). The former deals with the local creation of pieces of new knowledge

from existing pieces of knowledge; the latter is more oriented towards global optimization, and creates new knowledge by integrating large amounts of existing knowledge. These two algorithms themselves interact in several ways, representing the necessary interdependence of local and holistic cognition.

Architecturally, the Novamente system consists of a set of functionally specialized lobes, along the lines depicted in Figure 1 above. Each lobe contains a hypergraph representing declarative, procedural and episodic knowledge, and also contains a number of objects called MindAgents. Some of the MindAgents perform basic system maintanence operations (I/O, caching to disk, system statistics collection), but most of which contain cognitive algorithms applying PTL and BOA in conjunction with simple heuristics to carry out particular cognitive tasks like procedure learning, probabilistic inference on declarative knowledge, language parsing, and so forth. Figure 2 depicts the architecture of an individual Novamente lobe, and Figure 3 shows a collection of lobes networked together.

The "attention allocation" component of Novamente takes care of credit assignment, via using PTL and BOA to determine which Atoms and combinations of Atoms have been useful in which contexts in the past. This information is used to adjust the importance parameters of each Atom, which in turn determines how much attention the system's cognitive processes pay to each Atom. This leads to the formation of what are called "maps" – collections of Atoms that are habitually activated together, either all at once or in a particular habitual sequence. These maps can represent both declarative and procedural knowledge: they are an emergent level of knowledge representation, ensuing indirectly from Novamente's explicit Atom-based knowledge representation and its attention allocation dynamics. Table 5 reviews the basic types of Novamente maps.

Of course, the meaning of all these details lies in the integrated system behavior. Table 6 gives an indication of how Novamente, as a whole system, carries out various particular AI tasks differently than competing AI systems. But the crux of the matter, of course, is whether Novamente can ultimately lead to the emergence of the phenomenal self and a sense of will and self-awareness, qualitatively similar to what exists in the human mind. Unlike other AGI designs out there, it has been explicitly designed to do so. But the proof or disproof will be in the pudding.

The key question is whether Novamente, when placed in a simulated environment like AGI-SIM and interacted with by humans using linguistic and "physical" means, will develop a sophisticated and useful "self-map" – a map that models its own self in its interaction with the world and in its internal dynamics. This self-map need not be entirely accurate – no human's is – but it needs to be accurate enough to survive in the mind, and to be oriented toward particular tasks, thus creating a "focus of awareness" out of the "moving bubble of attention" supplied by Novamente's attention-allocation component.

Novamente Design Aspect	Primary Functions
Nodes	Nodes may symbolize entities in the external world, simple executable processes abstract concepts, or components in relationship-webs signifying complex concepts or procedures
Links	Links may be n-ary, and may link Nodes or other Links; they embody various types of relationships between concepts, percepts or actions. The network of Links is a web of relationships.
MindAgents	A MindAgent is a software object embodying a dynamic process such as activation spreading or first-order logical inference. It acts directly on individual Atoms, but is intended to induce and guide dynamic system-wide patterns.
Mind OS	The Mind OS builds on a distributed processing framework to enable distributed MindAgents to act efficiently on large populations of Nodes and Links
Maps	A Map represents declarative or procedural knowledge as a pattern of many Nodes and Links
Units	A Unit is a collection of Nodes, Links and MindAgents devoted to carrying out a particular function such as vision processing, language generation, or a specific information processing style such as highly-focused concentration

Table 2. Major Aspects of the Novamente AGI Design

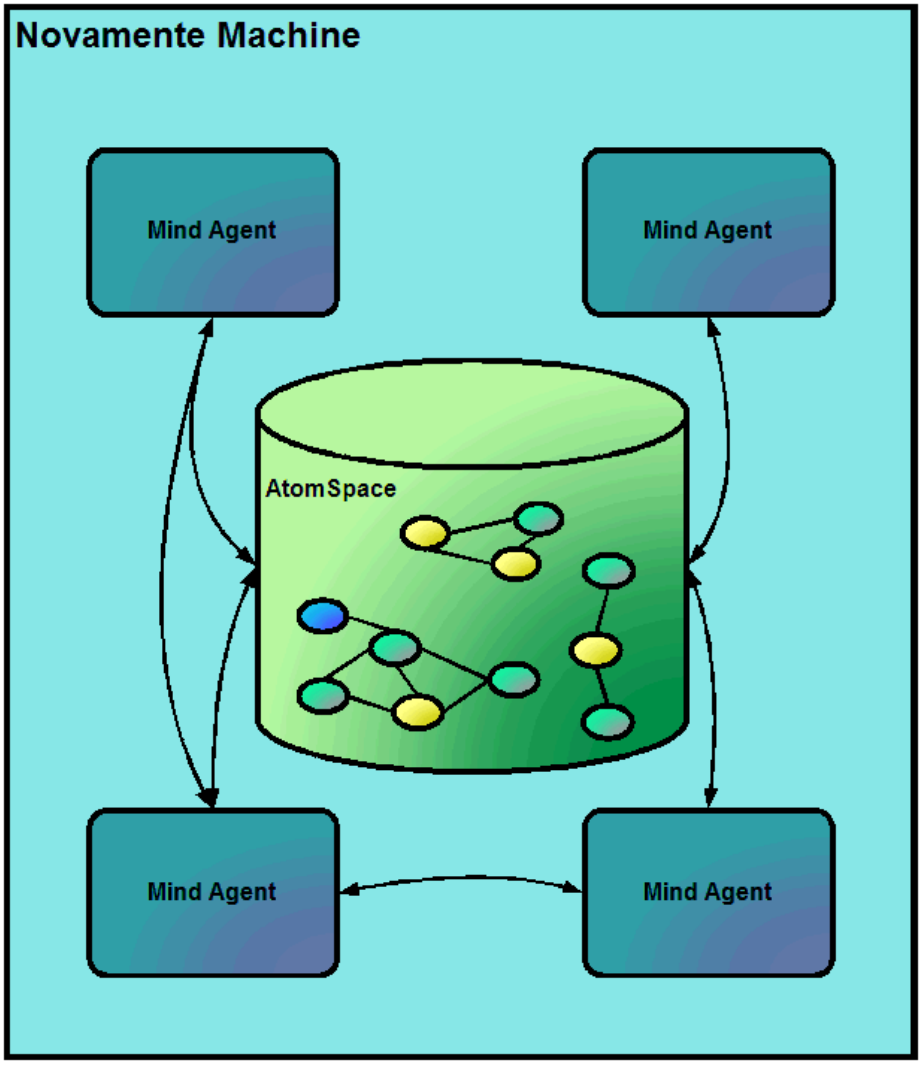

Figure 2. Conceptual Architecture of the Novamente "Mind OS" Layer

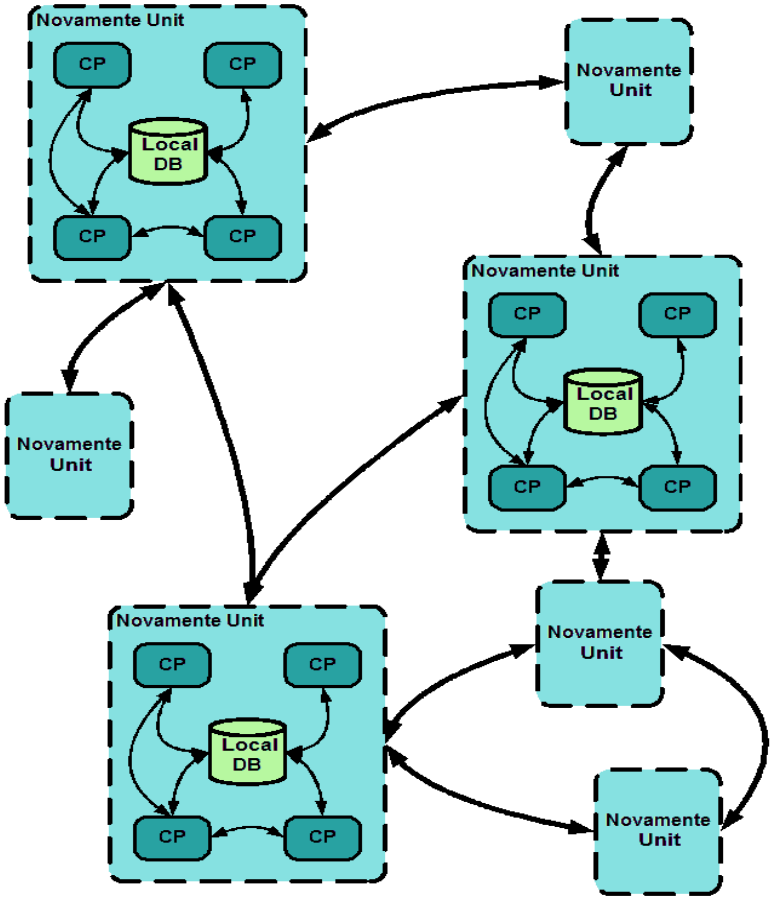

Figure 3. A Novamente Instance as a Distributed System

Node Variety	Description
Perceptual Nodes	These correspond to perceived items, like WordInstanceNode, CharacterInstanceNode, NumberInstanceNode, PixelInstanceNode
Procedure Nodes	These contain small programs called "schema,"[4] and are called SchemaNodes. Action Nodes that carry out logical evaluations are called PredicateNodes.
ConceptNodes	This is a "generic Node" used for two purposes. An individual ConceptNode may represent a category of Nodes. Or, a Map of ConceptNodes may represent a concept.
Psyche Nodes	These are GoalNodes and FeelingNodes, which are special PredicateNodes that play a special role in overall system control, in terms of monitoring system health, and orienting overall system behavior.

Table 3. Novamente Node Varieties

[4] Note that the use of the term "schema" inside Novamente derives directly from the use of this term in philosophy, e.g. the work of Immanuel Kant. It is different from the use of the term in contemporary database theory.

Link Variety	Description
Logical links	These represent symmetric or asymmetric logical relationships , either among Nodes (InheritanceLink, SimilarityLink), or among links and PredicateNodes (e.g. ImplicationLink, EquivalenceLink)
MemberLink	These denote fuzzy set membership
Associative links	These denote generic relatedness, including HebbianLink learned via Hebbian learning, and a simple AssociativeLink representing relationships derived from natural language or from databases.
ExecutionOutputLink	These indicate input-output relationships among SchemaNodes and PredicateNodes and their arguments
Action-Concept links	Called ExecutionLinks and EvaluationLinks, these form a conceptual record of the actions taken by SchemaNodes or PredicateNodes
ListLink and concatListLink	These represent internally-created or externally-observed lists, respectively

Table 4. Novamente Link Varieties

Map Type	Description
Concept map	a map consisting primarily of conceptual Nodes
Percept map	a map consisting primarily of perceptual Nodes, which arises habitually when the system is presented with environmental stimuli of a certain sort
Schema map	a distributed schema
Predicate map	a distributed predicate
Memory map	a map consisting largely of Nodes denoting specific entities (hence related via MemberLinks and their kin to more abstract Nodes) and their relationships
Concept-percept map	a map consisting primarily of perceptual and conceptual Nodes
Concept-schema map	a map consisting primarily of conceptual Nodes and SchemaNodes
Percept-concept-schema map	a map consisting substantially of perceptual, conceptual and SchemaNodes
Event map	a map containing many links denoting temporal relationships
Feeling map	a map containing FeelingNodes as a significant component
Goal map	a map containing GoalNodes as a significant component

Table 5. Example Novamente Map Types

MindAgent	Function	Development Status
First-Order Inference	Acts on first-order logical links, producing new logical links from old using the formulas of Probabilistic Term Logic	Complete
LogicalLinkMining	Creates logical links out of nonlogical links	Complete
Evolutionary Predicate Learning	Creates PredicateNodes containing predicates that predict membership in ConceptNodes	Complete
Clustering	Creates ConceptNodes representing clusters of existing ConceptNodes (thus enabling the cluster to be acted on, as a unified whole, by precise inference methods, as opposed to the less-accurate map-level dynamics)	Complete
Activation Spreading	Spreads activation among Atoms in the manner of a neural network	Complete
Importance Updating	Updates Atom "importance" variables and other related quantities	Implemented in prototype form
Concept Formation	Creates speculative, potentially interesting new ConceptNodes	Implemented in prototype form
Evolutionary Optimization	A "service" MindAgent, used for schema and predicate learning, and overall optimization of system parameters	Complete
Hebbian Association Formation	Builds and modifies HebbianLinks between Atoms, based on a PTL-derived Hebbian reinforcement learning rule	Implemented in prototype form
Evolutionary Schema Learning	Creates SchemaNodes that fulfill criteria, e.g. that are expected to satisfy given GoalNodes	Partially implemented
Higher-Order Inference	Carries out inference operations on logical links that point to links and/or PredicateNodes	Partially implemented
Logical Unification	Searches for Atoms that mutually satisfy a pair of PredicateNodes	Not yet implemented
Predicate/Schema Formation	Creates speculative, potentially interesting new SchemaNodes	Not yet implemented
Schema Execution	Enacts active SchemaNodes, allowing the system to carry out coordinated trains of action	Partially mplemented
Map Encapsulation	Scans the AtomTable for patterns and creates new Atoms embodying these patterns	Not yet implemented
Map Expansion	Takes schemata and predicates embodied in nodes, and expands them into multiple Nodes and links in the AtomTable (thus transforming complex Atoms into Maps of simple Atoms)	Not yet implemented
Homeostatic Parameter Adaptation	Applies evolutionary programming to adaptively tune the parameters of the system	Implemented in prototype form

Table 6. Primary Novamente MindAgents

Cognitive Task	Standard Approaches	Challenges	Novamente Approach
Logical Inference	Predicate, term, combinatory, fuzzy, probabilistic, nonmonotonic or paraconsistent logic	• Accurate management of uncertainty in a large-scale inference context • "Inference control": Intelligent, context-appropriate guidance of sequences of inferences	• Probabilistic Term Logic tuned for effective large-scale uncertainty management • Inference control carried out via a combination of inferential and noninferential cognitive processes
Attention Allocation	Blackboard systems, neural net activation spreading	The system must focus on user tasks when needed, but also possess the ability to spontaneously direct its own attention without being flighty or obsessive	Novamente's nonlinear, probabilistic inference based Importance Updating Function combines quantities derived from neural-net-like activation spreading and blackboard-system-like cognitive-utility analysis
Procedure Learning	Evolutionary programming, logic-based planning, feedforward neural networks, reinforcement learning	Techniques tend to be unacceptably inefficient except in very narrow domains	A synthesis of techniques allows each procedure to be learned in the context of a large number of other already-learned procedures, enhancing efficiency considerably
Pattern Mining	Apriori, genetic algorithms, logical inference, search algorithms	Finding complex patterns requires prohibitively inefficient searching through huge search spaces	Integrative cognition is designed to hone in on the specific subset of search space containing complex but compact and significant patterns
Human Language Processing	Numerous parsing algorithms and semantic mapping approaches: context-free grammars, unification grammars, link grammars; conceptual graphs, conceptual grammars…	Integrating semantic and pragmatic understanding into the syntax-analysis and production process	• Syntactic parsing is carried out via logical unification, in a manner that automatically incorporates probabilistic semantic and pragmatic knowledge. • Language generation is carried out in a similarly integrative way, via inferential generalization

Cognitive Task	Standard Approaches	Challenges	Novamente Approach
Self-Modeling (the creation of a "phenomenal self")	No current AI system or AGI design addresses this	• Creating a representational system sufficiently sophisticated to represent something as complex as a self in a compact way • Creating learning algorithms capable of the large-scale pattern-recognition prowess required to recognize something as large and abstract as a self in the large body of relevant but noisy data available to an embodied intelligence	This is viewed as a pattern recognition and inference problem similar to many others confronted by Novamente, but larger in scale. Novamente contains specific algorithms for mining patterns in its internal knowledge-hypergraph and explicitly embodying these patterns in new subgraphs. The self is one such pattern.

Table 7. Comparison of Approaches to Several Cognitive Tasks

4.4 Probabilistic Term Logic

In this section we will dig into one aspect of Novamente cognition in a moderate amount of detail: the Probabilistic Term Logic (PTL) inference module. The choice to expound on PTL in detail here, instead of some other aspect of Novamente cognition, is somewhat arbitrary, and is definitely based on pragmatic considerations – not because PTL is the most important part. Of the two key cognitive algorithms underlying Novamente, BOA is less original than PTL and is outlined in (Pelikan, 2002), although its application in a Novamente context involves many unique features (since in Novamente it acts on combinatory logic trees rather than Pelikan's bit strings). On the other hand Novamente's approach to attention allocation can't really be presented except in the context of PTL.

PTL is a highly flexible inference framework, applicable to many different situations, including inference involving uncertain, dynamic data and/or data of mixed type, and inference involving autonomous agents in complex environments. It was designed specifically for use in Novamente, yet also has applicability beyond the Novamente framework. PTL has been applied in several areas:

- To draw inferences regarding the relationships between concepts observed in news articles and messages, and extracted via language processing software

- To make guesses as the functional categories to which various little-understood genes belong, based on integrating quantitative gene expression data with data from biological ontologies
- To integrate the results of predictive models learned by BOA, predicting the density of vehicles in various regions of a map as it changes over time

Current research involves using PTL to reason on the output of a sophisticated natural language parser.

The goals motivating the development of PTL were the desire to have a practical, scalable inference system that operates consistently with probability theory. Although it uses probabilistic methods, PTL does not require a globally consistent probability model of the world, but is able to create locally consistent models of local contexts, and maintain a dynamically-almost-consistent overall world-model, dealing gracefully with inconsistencies as they occur. It encompasses both abstract, precise mathematical reasoning and more speculative hypothetical, inductive, and/or analogical reasoning; and it encompasses the inference of both declarative and procedural knowledge. It deals with inconsistent initial premises by dynamically iterating into a condition of "reasonable almost-consistency and reasonable compatibility with the premises", thus, for example, perceiving sensory reality in a way compatible with conceptual understanding, in the manner similar to that developed in the contemporary neural network literature, see e.g. (Haikonen 2003). Finally, it has the property that it makes most humanly simple inferences appear brief, compact and simple. For a sustained argument that term logic exceeds predicate logic in this regard, see (Sommers et al, 2000).

On a technical level, one difference between PTL and standard probabilistic inference frameworks is that PTL deals with multivariable truth values. Its minimal truth value object has two components: strength and weight of evidence. Another difference is PTL's awareness of context. Each PTL inference takes place in some context, which can be universal (everything the system has ever seen), local (only the information directly involved in a given inference), or many levels between.

PTL is divided into two portions: first-order and higher-order. First-order PTL deals with probabilistic inference on (asymmetric) inheritance and (symmetric) similarity relationships, where different Novamente link types are used to represent intensional versus extensional relationships (Wang, 1995). Example inference rules are deduction (A à B, B à C |- A à C), induction and abduction (shown in Figure 1), inversion (Bayes rule), similarity-to-inheritance-conversion, and revision (which merges different estimates of the truth value of the same atom).

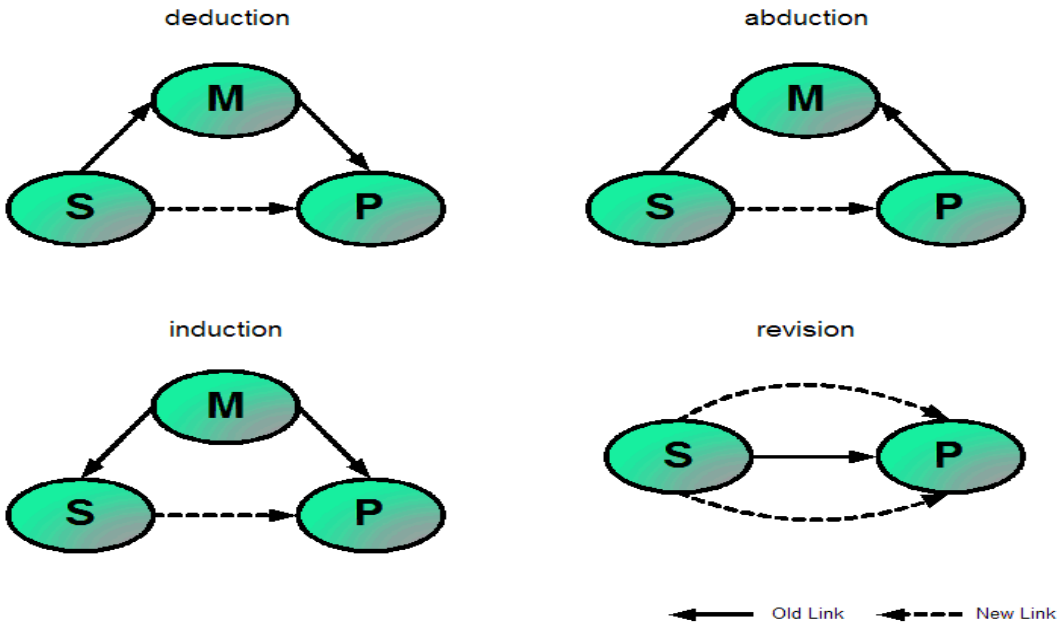

Figure 4. First-Order PTL Inference on InheritanceLinks

Each inference rule comes with its own quantitative truth value formula, derived using probability theory and related considerations. A simple quantitative example of first-order PTL deductive inference is:

```
InheritanceLink mud dangerous (.8,.7)
SimilarityLink sand mud (.6,.99)
|-
InheritanceLink sand dangerous (.31,.98)
```

The number-pairs such as (.8,.7) refer to the two components of a PTL truth value – the probability is .8, and the .7 represents the amount of evidence on which this probability estimate was based. The detailed trail of PTL inference in this case is as follows. First the similarity relation is converted to inheritance, yielding

```
InheritanceLink sand mud (0.38, 0.98)
```

Then the deduction

```
InheritanceLink sand mud (0.38, 0.98)
InheritanceLink mud dangerous (.8,.7)
|-
InheritanceLink sand dangerous (0.31,0.98)
```

is performed.[5] It's worth noting that the truth values in PTL combine both intensional and extensional information: the inheritance between mud and dangerous may be extensional, in the sense of deriving from actual observed instances of mud being dangerous; whereas the similarity between sand and mud is intensional, because it doesn't derive from there being a lot of instances that are both sand and mud, but rather from there being a lot of properties shared by sand and mud.

Higher-order PTL deals with inference on links that point to links rather than nodes, and on predicates and schemata. The truth value functions here are the same as in first-order PTL, but the interpretations of the functions are different. This aspect of PTL allows inference on complex patterns and procedures – an area in which alternate approaches to uncertain inference are extremely weak. We will review some examples of higher-order PTL below.

In sum, the PTL inference framework is a sophisticated, well-tuned approach to carrying out large-scale probabilistic inference in a real-world context. It has already been used for handling both quantitative, linguistic and mathematical inference, and is well suited to serve as a bridge between quantitative sensorial information and symbolic knowledge.

4.4.1 Intensional and Extensional Logical Relationships

One of the key concepts in PTL is the distinction between intensional and extensional logical relationships. The Subset relationship is what we call an extensional relationship – it relates two sets according to their members. The strength of the SubsetLink between A and B denotes the percentage of A's that are also B's. PTL handles Subset relationships but also deals with intensional relationships – relationships that relate sets according to the patterns that are associated with them.

Conceptually, the intension/extension distinction is very similar to that between a word's *denotation* and *connotation*. For instance, consider the concept "bachelor." The extension of "bachelor" is typically taken to be *all and only the bachelors in the world* (a very large set). In practical terms, it means all bachelors that are known to a given reasoning system, or specifically hypothesized by that system. On the other hand, the *intension* of "bachelor" is the set of properties of "bachelor," including principally the property of being a *man*, and the property of being *unmarried*.

Some theorists would have it that the intension of "bachelor" consists solely of these two properties, which are "necessary and sufficient conditions" for bachelorhood; PTL's notion of intension is more flexible, it may include necessary and sufficient conditions but also other properties, such as the fact that most bachelors have legs, that they frequently eat in restaurants, etc. These other properties allow us to understand how the concept of "bachelor" might be stretched in some contexts – for instance, if one read the sentence "Jane Smith was a more of a bachelor than any of the men in her apartment building," one could make a lot more sense of it using the concept "bachelor"'s full PTL

---

5 PTL requires "node probabilities" for this inference, which are defined relative to a relevant context. The example uses the values mud: 0.001, sand: 0.05, dangerous: 0.015.

intension, than one could make using only the necessary-and-sufficient-condition intension.

To understand the relation between intensional and extensional inheritance (Subset) in practice, consider the example of fish and whales. Extensionally whales are not fish, i.e.

```
Subset whale fish <.0001>
```

But intensionally, the two share a lot of properties, so we may say perhaps

```
IntensionalInheritance whale fish <.7>
```

The essential idea underlying PTL's treatment of intension is to associate both fish and whale with sets of patterns – $fish_{PAT}$ and $whale_{PAT}$, the sets of patterns associated with fish and whales. We then interpret

```
IntensionalInheritance whale fish <.7>
```

as

$$\texttt{Subset } whale_{PAT}\ fish_{PAT}$$

And we then define Inheritance proper as the disjunction of intensional and extensional (subset) inheritance, i.e.

```
Inheritance A B
```

is defined as

```
OR
      Subset A B
      IntensionalInheritance A B
```

Why do we think intensional relationships are worth introducing into PTL? This is a cognitive science rather than mathematical question. We hypothesize that most human inference is done not using subset relationships, but rather using composite Inheritance relationships. And, consistent with this claim, we suggest that, in most cases, the natural language relation "is a" should be interpreted as an Inheritance relation between individuals and sets of individuals, or between sets of individuals – not as a Subset relationship. For instance,

```
"Fluffy is a cat"
```

as conventionally interpreted is a combination extensional/intensional statement, as is

```
"Cats are animals."
```

This statement means not only that examples of cats are examples of animals, but also that patterns in cats tend to be patterns in animals.

Philosophically, one may ask why a pattern-based approach to intensional inference makes sense. Why isn't straightforward probability theory enough? The problem is – to wax poetic for a moment -- that the world we live in is a special place, and accurately reasoning about it requires making special assumptions that are very difficult and computationally expensive to explicitly encode into probability theory. One special aspect of our world is what Charles Peirce referred to as "the tendency to take habits": the fact that "patterns tend to spread," i.e. if two things are somehow related to each other, the odds are that there are a bunch of other patterns relating the two things. To encode this tendency observed by Peirce in probabilistic reasoning one must calculate P(A|B) in each case based on looking at the number of other conditional probabilities that are related to it via various patterns. But this is exactly what intensional inference, as defined in PTL, does. This philosophical explanation may seem somewhat abstruse – until one realizes how closely it ties in with human commonsense inference, and with the notion of inheritance as utilized in natural language.

4.4.2 Analogical Reasoning in PTL

As an illustration of how PTL works in practice, we'll now consider a simple example of analogical inference. Imagine Novamente is playing detective, and is hunting for an individual names Smith. Consider the following min-scenario: "*Novamente knows what kind of car Smith has, and asks other agents in the area if they've seen a similar car. They say they have seen it parked in a certain particular garage. The agent checks out that garage, but Smith's car isn't there. So it decides to check out nearby garages.*"

We now explain in detail how PTL reasons that, because people often hide in houses nearby their houses, perhaps cars are often hidden in garages nearby the garages they usually are parked in.

First we introduce some notation, informally. One piece of knowledge it needs to carry out this inference is that

```
person hide_in $X implies $X is house (.2,.9)
```

This is a PTL probabilistic implication relationship. In the notation used standardly to describe Novamente nodes and links, this implication would be written as:

```
ImplicationLink (.2,.9)
      EvaluationLink hide_in
            ListLink (person, $X)
      InheritanceLink $X house
```

Here *person* and *house* are ConceptNodes and *hide_in* is a PredicateNode. Note that these ConceptNodes and PredicateNodes are merely tokens signifying the network of nodes and links within Novamente that embody the concepts of hiding, people, and so

forth. The overall network of nodes and links signifying hiding may be quite large and complex, embodying a knowledge of many properties and instances of hiding, but nevertheless there may also be a single ConceptNode linked to the PhraseNode *hide_in*, which has the property that it's activated if and only if the overall "map" of nodes and links signifying hiding-in is activated.

The system may also know that people tend to hide in houses near the ones they live in (an implication learned by the Novamente predicate mining heuristic)

```
AND ($X is person, $Y is house, $Z is house, $X live_in $Y, $X hide_in $Z)
                    implies $Y is_near $Z (.4,.7)
```

And it may also know

```
house is building <.9,.99>
```

from which it can reason (via PTL deduction)

```
AND($X is person, $Y is building, $Z is building, $X live_in $Y, $X hide_in $Z)
                    implies $Y is_near $Z (.11,.87)
```

Now if it knows that people are similar to cars in some regards, i.e.

```
person is like car (.2,.99)
```

(a SimilarityLink relationship) then it can conclude that

```
AND($X is car,$Y is building, $Z is building, $X live_in $Y, $X hide_in $Z)
                  implies $Y is_near $Z (.05, .23)
```

If it knows that

```
garage is building (.99,.99)
car regularly_park_in garage (.9,.9)
regularly_park_in is like live_in (.4,.8)
```

it can conclude that[6]

```
AND($X is car,$Y is garage, $Z is garage, $X regularly_park_in $Y, $X hide_in $Z)
                    implies $Y is_near $Z (0.032, .23)
```

In cognitive science terms, this general conclusion was arrived at via analogy to the original observation about people hiding in houses. In PTL terms, the single "analogy" step is decomposed into a number of probabilistic inference steps. Note that the conclusion is less certain than the premise, because a number of somewhat shaky assumptions were used along the way (for instance, the similarity between people and cars, the similarity between parking and living).

---

6 In carrying out these inference we have assumed the following contextually appropriate node probabilities: person: .05, vehicle: .03, house: .02, building: .035, garage: .045, regularly_park_in 0.06, live_in 0.055

So, suppose there is a particular car that our autonomous cars is looking for, and this car has been seen parking in a certain garage several times. Then it may wish to make use of a rearrangement of the above, namely

```
AND($X is car,$Y is garage,$Z is garage,
 $X regularly_park_in $Y,$Y is_near $Z)
      implies $X hide_in $Z (s,N)
```

In order to derive the truth value of this, PTL needs to have an estimate of how many garages are near the garage the car regularly parks in. This of course will depend upon the particular garage in question. In a crowded part of the city, this number may be sufficiently large that the truth value strength s of the above statement is very low, so the car decides it's not worth searching through the nearby garages due to lack of time. On the other hand, if there aren't many garages around, then s will be reasonably large for any particular garage, so that spending a limited time searching in nearby garages is worthwhile.

5 Novamente vs. the Human Brain/Mind: Memory, Learning and Perception

Now, having explained a bit about Novamente and its conceptual underpinnings, we will review some specific relationships between the lessons of modern cognitive science, the SMEPH approach to mathematical mind-modeling, and the Novamente AI design.

Table 8 gives some rough intuitive correspondences between Novamente structures and human brain structures. This table should be taken with several grains of salt – clearly, the brain is not sufficiently well understood for a table like this to be made with a high degree of confidence. But we have our own intuitions based on the current state of knowledge, and we feel these are worth sharing.

Table 8 notwithstanding, we will mostly deal here with cognitive science proper rather than cognitive neuroscience, and introduce neuroscience ideas in cognitive context. As noted above, due to the incompleteness of our knowledge of the human brain, knowledge from brain science is really only useful for AGI design when it is coupled with knowledge from cognitive science. For instance, if it weren't evident psychologically that visual percepts are perceived, remembered and reasoned on as wholes, then the fact that a single visual percept typically activates widely distributed neurons might seem to have a quite different significance than the currently accepted one. Because of this cognitive fact we have the binding problem, and we have the interesting neurocognitive hypothesis that percepts are represented as attractors – a hypothesis with powerful AGI implications. But this AGI hypothesis would never have come out of the neuroscience in itself, it had to come from a neuro/cognitive understanding.

On the other hand, cognitive science results can be quite helpful for AGI in themselves, quite apart from whether they are backed up by any neuroscience knowledge.

An example is the cognitive science of abstract reasoning. Of course, though, cognitive science results are most interesting when backed up by brain science as well; and we will draw our greatest inspiration from those cases where the two disciplines coincide.

Human Brain Structure/Phenomenon	Primary Functions	Novamente Structure/Phenomena
Neurons	Impulse-conducting cells, whose electrical activity is a key part of brain activity	No direct correlate: Novamente's implementation level is different
Neuronal groups	Collections of tightly interconnected neurons, often numbering 10,000-50,000	Novamente nodes
Synapses	The junction across which a nerve impulse passes from one neuron to another; may be excitatory or inhibitory	Novamente links are like bundles of synapses joining neuronal groups
Synaptic Modification	Chemical dynamics that adapt the conductance of synapses based on experience; thought to be the basis of learning	The HebbianLearning MindAgent is a direct correlate. Other cognitive MindAgents (e.g. inference) may correspond to high-level patterns of synaptic modification
Dendritic Growth	Adaptive growth of new connections between neurons in a mature brain	Analogous to some heuristics in the ConceptFormation MindAgent
Neural attractors	Collections of neurons and/or neuronal groups that tend to be simultaneously active	Maps, e.g. concept and percept maps
Neural input/output maps	Composites of neuronal groups, mapping percepts into actions in a context-appropriate way	Schema maps
"Neural Darwinist" map evolution	Creates new, context-appropriate maps	Schema learning via reinforcement learning, inference, evolution
Cerebrum	Perception, cognition, emotion	The majority of Units in a Novamente configuration
Specialized cerebral regions (Broca's area, temporal lobe, visual cortex,…)	Diverse functions such as language processing, visual processing, temporal information processing,…	Functionally-specialized Novamente Units
Cerebellum	Movement control, information integration	Action-oriented units, full of action schema-maps
Midbrain	Relays and translates information from all of the senses, except smell, to higher levels in the brain	Schemata mapping perceptual Atoms into cognitive Atoms
Hypothalamus	(regulation of basic biological drives and controls autonormic functions such as hunger, thirst, and body temperature)	HomeostaticParameterAdaptation MindAgent, built-in GoalNodes
Limbic System	(control emotion, motivation, and memory)	FeelingNodes and GoalNodes, and associated maps

Table 8. Novamente vs. the Human Brain

Another preliminary note to be made is that the SMEPH approach doesn't necessarily break the mind down into components in the same ways as the mainstream of

modern cognitive science. For instance, memory and reasoning are typically considered as separate things, in the course of cognitive science research. Yet, in the SMEPH approach, it is considered that most acts of "memory retrieval" are actually coordinated acts of reasoning, "constructing" memories from stored knowledge. Similarly, reasoning and perceptual pattern recognition are typically considered as different things, yet in the SMEPH approach, perceptual pattern recognition is done via the same probabilistic equations used for abstract reasoning, deployed in simpler and more scalable ways.

These differences don't make it impossible to draw mappings between Novamente and the human mind/brain, but they do mean that these mappings must be drawn with care. In every case we've explored so far, when one probes deeply, one finds that the SMEPH/Novamente approach is harmonious with the ideas of some significant subset of cognitive science researchers. For instance (Riegler, 2005) advocates the constructive nature of memory, whereas (Goldstone et al, 2005) discusses parallels between perceptual learning and abstract cognition; etc. In many cases cognitive science divides mental process into categories based on convention and convenience; and when building an AGI one is confronted with a different notion of convenience – a division (like memory vs. reasoning) that's convenient for guiding the design of experiments on human subjects can be extremely inconvenient from the perspective of AGI design. In this regard Novamente is perhaps closer to neuroscience than cognitive science – neuroscientists are continually discovering feedback loops and dynamical and structural complexities that break through the simplistic divisions favored by many cognitive theorists. This is because neuroscientists, like AGI designers and engineers, are dealing with the necessary messiness of real complex systems, rather than with simplified theoretical abstractions.

5.1 Novamente's Memory vs. Human Memory

One area where Novamente clearly accords with cognitive science ideas is the division of memory into various subcomponents. The distinction between procedural, episodic and declarative memory is well-demonstrated both psychologically and neuroscientifically (Baddeley, 1999), and it is also quite natural in terms of Novamente's hypergraph knowledge representation. In fact the distinction between procedural and declarative knowledge already exists at the SMEPH level of abstraction, in the form of the distinction between Concept edges and Schema edges.

Declarative knowledge is naturally represented in Novamente via probabilistic-logical link types, whereas procedural knowledge is naturally represented using links explicitly representing actions taken. Episodic memory, finally, is naturally represented via links joining probabilistic-logical relationships defining sequences of events with records stored in an "experience database." The distinction between the three memory types becomes, in Novamente terms, a matter of representational efficiency.

Experiential episodes can be stored in declarative logical terms but this is extremely inefficient; so for practical purposes it's better to store experiences in another, less flexible form, and map only their high-level structure in declarative form. As the mind and its goals change, the experience database will be repeatedly revisited and its

contents re-represented declaratively in different ways based on different acts of pattern recognition.

On the other hand, procedural knowledge can also be stored in declarative form, and this is useful when one wants to reason about procedures. But it is generally the case that the most natural form of a procedure from the point of view of reasoning about the procedure is not the most efficient form from the point of view of actually executing the procedure. So for practical reasons one winds up with dual representations of procedures: an executable form that is compacted for execution efficiency, and an expanded form that's suitable for generalization and inference. A difference between AGI systems and the human brain comes up here: for humans it can be extremely difficult to create declarative forms for procedural knowledge. On the other hand, detail-level introspection is much easier for a software program than for a human brain, and procedural-to-declarative conversion doesn't need to be so problematic. This is one among many areas in which a slavish adherence to human neuropsychology is probably not clever AGI-design-wise.

Similarly, the distinction between short-term and long-term memory, fundamental to human psychology and well validated via neuroscience also turns out to be fundamental to AGI design, again for a combination of efficiency reasons and general SMEPH reasons. The "short-term memory" concept has undergone a number of name changes as cognitive theory has developed, but the basic idea is simple and solid. In a situation of limited processing power, not everything can be attended at once. The Novamente approach involves the assignment of "importance" parameters to nodes and links, and the use of inference and pattern recognition to update these importance values dynamically. Depending on the parameter values of the process, this dynamic updating will frequently lead to a situation in which a small percentage of nodes and links garner a vast majority of attention.

This overall process exists in SMEPH generally, not just in Novamente. In a SMEPH hypergraph, the edges and vertices that have been most useful will get more attention, but due to the nature of cognitive dynamics, there will often be groups of edges and vertices that, in a specific context, are particularly useful together as a unit.

In the Novamente design this dynamic phenomenon (the emergence of STM) is structurally reified via the use of a separate lobe (the AttentionalFocus) for the most important nodes and links. This seems qualitatively similar to how, in the human brain, the focus of attention is structurally reified via separate brain structures devoted to STM in its various forms.

The refinements of the notion of STM that have occurred over the last few decades are also well reflected in the Novamente/SMEPH approach. One recent discovery is that, prior to the registration of sensations in the STM, there is a preliminary process in which fleeting connections are drawn between sensations and knowledge stored in LTM. This emerges naturally from the framework in which there is one overall knowledge-network and the "STM" is simply the "moving activation bubble" of most-important-entities.

Another discovery that has become gradually solidified since Baddeley (1989) first systematically presented it is the existence of multiple modality-specific STM's, such as a visual-perception STM, a linguistic STM, and a generic "mental workspace."

This emerges naturally in the SMEPH approach, because the specialized schemata that handle a process like visual perception or language processing will naturally gather separate bubbles of highly-active nodes and links around them. Again this dynamic process is reified in the Novamente architecture via the creation of specific lobes for modality-specific AttentionalFocus.

In cognitive science terms, SMEPH leads to a flexible model of attention, as opposed to the more rigid filtering or late-selection based approaches that were popular among theorists in the past (Underwood, 1993). For instance, in the case where information comes in from two sensory channels and there are not enough resources to process both information-streams simultaneously, one achieves the result that: the earlier the stage at which selection between channels is possible (i.e. the closer the attentional focus is to the sensory level), the faster and more efficient the response to the attended channel, and the less is processed in the unattended channel. This accords with results from human psychology; for instance research on the timing of accessing word meanings (Luck, Vogel and Shapiro, 1996).

The "binding problem" that is so critical in modern cognitive science is less critical in the SMEPH domain. Thinking about binding in the context of AGI impels one to decompose the problem into two parts: a conceptual part and a physiological part. Conceptually, there is a question of the logic and the cognitive dynamics by which disparate percepts are bound into a unified whole. Then, physiologically, there is the question of how the brain executes this logic and dynamics, which is a subtle issue because the parts of the brain representing different parts of a unified percept are often physically widely distributed. AGI systems give rise to the conceptual problem but not the physiological one. The solution to the conceptual problem seems straightforward in the SMEPH context; it follows very closely Walter Freeman's (2001) ideas regarding the emergence of attractors in the brain. The linkages between sensation and LTM cause the *relationships between* the nodes and links involved in percepts corresponding to parts of a unified object to become important, which encourages the formation of predicates and concepts binding all of them together. From this perspective, the physiological "binding problem" then comes down to how the formation of dynamic associational and logical linkages occurs between percepts and concepts represented in distant brain regions – a very important question for neuroscience, but not directly relevant to AGI systems, except those that seek to emulate the way the brain uses three-dimensional geometry to help represent knowledge and guide cognitive and perceptual dynamics.

The nature of forgetting in human memory seems to an extent to represent general principles that are also applicable to SMEPH based systems. It has recently become clear that a substantial amount of the forgetting that occurs in the human mind can be attributed to memory interference rather than simply "running out of space" (Wixted, 2004). This sort of phenomenon occurs naturally as a consequence of importance dynamics: if two pieces of knowledge contradict each other then when one gets attention the other will tend not to, and the loser in the rivalry will gradually get its importance downgraded until it's forgotten. In this view, running out of space is the ultimate reason for forgetting, but interference can explicitly cause memory items to be deprioritized. Also, the known fact that humans rarely truly forget anything that's been fully learned (Baddeley, 1989) is reflected in the notion of long-term importance. In Novamente,

when an item is important on enough occasions, it achieves a high long-term importance, which basically guarantees that it will be preserved in the deep memory store (i.e. saved to disk) rather than being permanently forgotten.

Finally, regarding the particular structure of the contents of memory, cognitive science doesn't have anything definitive to say at the moment. The SMEPH model is reminiscent of semantic-network-based models of human memory, which originated with Quillian and have played a major role in many subsequent cognitive modeling approaches such as Anderson's work on ACT-R (1997). However, the nonlinear self-organizing dynamics in SMEPH is also reminiscent of Walter Freeman style theories in which knowledge is represented in dynamical attractors. At the moment semantic-network-style models seem best able to deal with the human mind's treatment of abstract, linguistic or mathematical declarative knowledge, whereas attractor-style models seem better able to deal with knowledge directly related to perception and action, and also to issues such as binding and the creation of unified percepts and unified phenomenal selves from conceptually and physically disparate components. The SMEPH approach unifies these two approaches by proposing a framework in which the importance levels of nodes and links in a semantic network display complex dynamics with attractors, and specific semantic network nodes in many cases "key" specific attractors.

5.2 Learning in Humans and in Novamente

"Learning theory," in psychology, began in earnest with behaviorist studies of rote learning by pigeons, dogs and other animals. Commonalities between learning behavior in humans and these other animals were correctly observed, and some basic principles of learning were enounced: contiguity (spatiotemporally nearby things are associated), frequency (conditional probabilities are tabulated), contingency and blocking (e.g. after it's learned that a light predicts a shock, if a tone is introduced every time the light occurs, there is inhibition against learning that the tone predicts a shock). This kind of behaviorist learning has been shown to roughly obey probabilistic principles; e.g. the foraging behavior of wild birds automatically adapts itself to constitute a near-optimal solution to the relevant multi-armed bandit problem (Alexander, 1996).

Neurally, behaviorist-style learning ties in naturally with Hebbian learning – an old hypothesis which is increasingly substantiated by neurological research. We now know a fair bit about the chemical, genomic and proteomic dynamics underlying the processes of neuronal long-term potentiation that implement approximations of Hebb's basic learning rule (to increase the conductance of synapses that are repeatedly used).

However, the connection between neuron-level Hebbian learning and organism-level behavior learning is not quite so direct as many naively believed in the past. To achieve animal-level behaviorist learning via a Hebbian simulated neural network can require quite complex dynamics in a neural net of substantial size; see e.g. (Wilson, 2000) which uses a sophisticated Hebbian-style neural network model, implemented in terms of continuous-valued neurons and differential equations, to simulate behavior learning in Siamese fighting fish. This sort of work also indicates the amount of subtle tuning of Hebbian learning that is necessary to get it to give meaningful and useful results. It is interesting to contrast Wilson's work with the refinements of the Hebbian

approach presented in Sutton and Barto's (1998) classic text on reinforcement learning. The former refines the basic Hebbian idea based on biological plausibility and quality of simulating biological learning behavior; the latter based on mathematical elegance and learning performance on computer science test problems. The clear message is that there are a lot of ways to tweak this basic learning mechanism, and the best way to tweak it for a given purpose is not at all obvious. In particular, we have little idea at present how Hebbian learning would be modified and adjusted to give rise to the type of intermediate-level learning we see in the human brain.

Clearly there is something powerful and valuable in the idea of Hebbian learning –something AGI designers should not ignore. However, from an AGI point of view, one is led to wonder whether it is sensible to implement Hebbian learning directly, or to try to figure out what higher-level learning dynamics neural Hebbian learning is giving rise to, and emulate *these* in software. In this regard the close connections between Hebbian learning and probability theory (Sutton and Barto, 1997) are highly interesting.

In (Goertzel, 2003), it is argued that Hebbian learning on the neuron level naturally gives rise to probabilistic reasoning on the level of neuronal clusters or sets of neuronal clusters. This suggests that if one hypothetically associates nodes in a semantic hypergraph with neuronal clusters or sets thereof, one can associate neural Hebbian learning with probabilistic inference on the probabilistic link weights in the hypergraph. I.e., it suggests that perhaps the reason Hebbian learning works so well in the brain is that it gives rise to approximate probabilistic inference on the level of the semantic hypergraph emergent from the brain. Of course, this is a speculative theory – cognitive neuroscience hasn't yet informed us how semantic-hypergraph-nodes are grounded in the brain, so we can't even ask detailed questions about how various sorts of inferences about their interrelationships are neurally grounded. But it is a speculation that accords with all available evidence, and is intuitively harmonious with existing knowledge in neuroscience, cognitive science and AI.

Hebbian learning is, on the face of it, a highly "local" learning method: it works by making incremental modifications to existing neurally based knowledge. In computer science local learning methods are valuable, but there's also a place for more global methods, which make big leaps to find answers far removed from existing knowledge. One of the most powerful global learning methods available is evolutionary learning, which takes many guises including genetic algorithms (Holland, 1992) and genetic programming (Koza, 1992), and roughly emulates the process of evolution by natural selection. It is known that the immune system adapts to new threats via a form of evolutionary learning, and Edelman (1987) has proposed that the brain does as well, evolving new "neuronal maps" – patterns of neural connection and activity spanning numerous neuronal clusters – that are highly "fit" in the sense of contributing usefully to system goals. He and his colleagues have run computer simulations showing that Hebbian-like neuronal dynamics, if properly tuned, can give rise to evolution-like dynamics on the neuronal map level ("neuronal group selection"). This is very interesting, and is something that could potentially be implemented in a SMEPH context as well, where one could see emergent evolutionary learning arise from link-level probabilistic inference.

Novamente involves this sort of "emergent evolutionary map-level dynamics" phenomenon but, in a significant design decision, it also involves explicit evolutionary programming using BOA (Pelikan, 2002), an algorithm related to but more efficient than genetic programming. Like the creation of specific lobes for attentional focus and sensory modality focus, this is a case of explicitly choosing architectural features to match harmoniously with emergent phenomena.

So, Novamente's two key learning algorithms – probabilistic inference via PTL and evolutionary learning via BOA – may be seen to correspond to two key aspects of learning in the brain: Hebbian learning and neuronal group selection. And as with their neural correlates, these two learning algorithms fit naturally together – but for moderately different reasons. BOA and PTL fit together via their mutual reliance on probability theory, whereas Hebbian learning and neuronal group selection fit together because of their common reliance on the physiology and electrochemistry of neurons and other brain cells.

5.3 Reasoning in Humans and Novamente

Logical reasoning is often taken as the most uniquely human cognitive characteristic. Language is the most unique human faculty in everyday terms, but logic is in a sense the apex of human achievement. It's logical reasoning that's led us to modern civilization – to such things as mathematics, science, and the institutions of democratic governance. Furthermore, inference is closely related to the issue of consciousness and awareness, since in humans conscious supervision and control of thought seems to be necessary for confronting a new problem with logical inference tools.

Logical reasoning is also something that is fairly close in some ways to the internal operations of computers, so it's not surprising that a lot of AI research has focused on the area of reasoning. However, automated reasoning systems have performed very poorly to date, unable to carry out either

- everyday commonsense reasoning in the manner of a small child (though this was the goal of Cyc from the start, after a couple decades Cyc is still nowhere near achieving this goal)
- mathematical theorem-proving without detailed human guidance, beyond the level of simple theorems in set theory (Robinson and Voronkov, 2001)

The reason for this poor performance is moderately subtle. It's not that the AI programs are using bad reasoning rules. Their reasoning rules are correct, in fact more so than the reasoning rules implicitly used by humans in many cases. Humans are prone to stupid reasoning errors (Pietelli-Palmarini, 1996), and this often harms us in practical situations. Rather, the problem is that AI systems are applying the reasoning rules to the wrong sorts of entities, and they don't understand in what order to apply their reasoning rules – how to design a contextually-appropriate inference trajectory.

Regarding the "wrong sorts of entities" problem, there is a large literature on the nature of human concepts. This pertains to the intension/extension distinction discussed

above in a PTL context. The classical school of thought held that concepts were defined by necessary and sufficient condition , but this has largely given way to a theory holding that concepts are defined mainly by prototypes and exemplars (Hunt and Ellis, 1999). Novamente and SMEPH suggest that both of these theories have some truth to them – that both necessary/sufficient conditions and prototypes/exemplars may be considered as probabilistic relationships in a concept hypergraph. Novamente theory also suggests a third aspect to the definition of concepts, which may be important for human as well as AI cognition: pattern-based intensional definition, wherein a concept is defined partly by the set of patterns associated with it.

Humans seem to have good heuristics for figuring out which concepts to use to describe a given situation. For instance, when perceiving a watch, a human must decide whether to think about it as a "watch", an "object", a "self-winding ladies' analog wristwatch", etc. The choice depends on context – i.e., it depends on which classification is going to be most useful for the inferences one wants to draw. One heuristic humans often use is that objects that are atypical of basic level objects tend to be named and identified at a subordinate level (Jolicoeur, Gluck and Kosslyn, 1984). In general, this "most useful level of categorization" problem is a subcase of the overall inference control problem – the problem of knowing which inference steps to carry out in which order.

The SMEPH/Novamente approach to this issue is conceptually simple: inference control strategies are represented as schema, and may be learned just like any other kind of procedural knowledge. The trick is that this is a very difficult learning problem. And this is where experiential learning comes in: humans learn to reason by starting out in very simple situations (cf. Piaget et al, 2001), and once their inference control strategies are thus honed, they are ready to deal with slightly more complex situations, etc. It may be that in order to learn to reason effectively, an AI must go through the same sort of series of steps. Current Novamente applications have sidestepped this problem by using narrow-AI-style, hand-coded, purpose-specific inference control strategies.

5.4 Human Language Processing

Psycholinguistics is a burgeoning field (Gleason, 2004), yet is plagued by basic unanswered questions. There is still no real consensus on how much of human language acquisition is based on pure experiential learning, and how much is based on learned adaptations and parameter-tunings to genetically provided modules. Attempts to have computer programs learn language based on pure statistical analysis of text have run into a brick wall far short of human-level language comprehension (Manning and Schutze, 1999), yet this doesn't tell us much about the learnability of language by an embodied system. Some early work on "symbol grounding" (e.g. learning word meanings via correlating word usage with robot sensor input) has been carried out in recent years (Roy and Mukherjee, 2005), but no one has yet carried out experiments in the area of experiential grounding of complex syntactic or semantic relationships.

Human language processing is typically called NLP or "natural language processing", but of course human language is not necessarily natural for a nonhuman intelligence. Learning human language without a human body or human evolutionary

heritage, is a much harder problem than learning human language in the possession of such endowments. In the case of Novamente we have designed a special communication format called Psynese, which is unlike any human language and is a form of linguistic communication that's much more compatible with Novamente's nature. We have also toyed with the idea of building an interface to Novamente using Lojban (Nicholas and Cowan, 2003), a constructed language that has an unambiguous syntax and a foundation in predicate logic yet is also suitable for informal human conversation. But, in spite of the unnaturalness of human language for Novamente, there is no question that giving the system human language processing ability is a must. Novamentes and humans have too much to learn from each other. Novamente must eventually be able to read human-written research papers, explain its ideas to humans in ways they can understand, obtain moral and pragmatic guidance from humans, and develop ideas collaboratively with humans through conversations.

The brick wall hit by non-embodied computational language learning algorithms takes various forms, but it always somehow comes down to one thing: the dependency of language on subtle interdependencies between syntax, semantics and pragmatics. And this dependency has everything to do with the deep link between embodied experience and advanced cognition.

In fact, we are intimately familiar with these issues because we have experienced them ourselves! In parallel with developing Novamente as a would-be AGI system, we have also spent a great deal of time in the last few years building specialized commercial software applications based on the Novamente platform. This has been done out of financial necessity rather than out of a believe that narrow AI is the best path to general AI, but in some cases it has led to work that's been valuable for general AI as well as for short-term narrow AI applications. Among our Novamente-based narrow-AI projects has been one called INLINK (Goertzel et al, 2005), which is natural-language-focused. Rather than being a typical, purely unsupervised natural language understanding system, INLINK is an interactive system, in which the user types in natural language sentences and then interacts with the user interface to be sure that the system has made the correct interpretation of the sentence. Because we needed to make INLINK work in a reasonably short time-frame, we couldn't implement language understanding in the way we really wanted to – based on experiential learning, grounding of linguistic terms in an embodied world, and so forth. Instead we implemented a specialized, hard-coded NLP system that handles syntax processing and feeds its outputs into the Novamente Atomspace for subsequent semantic analysis.

Experimenting with the INLINK system has been interesting in terms of seeing exactly how far a very cleverly constructed rule-based NLP system can be pushed. We knew we would eventually run up against a brick wall, a fundamental limitation in the system's capability due to the fact that its syntax processing was being done without deep feedback from the semantic layer. But we didn't know exactly where this limit would be. We have found that it arises most clearly in the issue of grounding prepositions and other function words. The subtle meanings of these little words like "of", "to", "by" and "with" are not well captured in dictionary definitions; we have created our own "Preposition Wordnet" dictionary that captures them better than other resources, but it's still far from adequate. It seems qualitatively clear to us that in order to really manipulate

the semantics of these little words accurately, some experiential grounding in a collaborative world-environment will be necessary. The issue is not so much syntax parsing as the mapping of syntactic output into semantically meaningful relationships suitable for guiding reasoning.

5.4.1 Example of Semantic Analysis

To more concretely illustrate the issues involved with syntax parsing and prepositional semantics and inference, we will show here three representations for the three almost-semantically-identical sentences, produced by INLINK's semantic analysis component. We will discuss what INLINK does and what its shortcomings are, as compared to the requirements posed by AGI.

For sake of compactness, in these examples we show only a limited subset of the actual Novamente links created in processing the sentences. We omit WSLinks (which link WordNodes to ConceptNodes) and the like, and show only semantically meaningful links between ConceptNodes. ConceptNodes are denoted by the names of the WordNodes most closely linked to them, and other nodes such as those denoting tense (e.g. %pres_ongoing) are denoted by intuitive shorthand names

Finally, in these examples, links are shown in a relational-logic style, where the notation R(X,Y) is used both for Novamente link types R and for predicates R, i.e. it may mean either

- that a link of type R exists between the node or link X and the node or link Y, or
- that an Evaluation link exists between the predicate R and the List Atom (X, Y)

For the present purposes this distinction is not an important one (though it is important for Novamente dynamics). Recall also that in Novamente links may span links as well as nodes.

Without further ado, the three examples:

Amir is a friend of James.

```
Inheritance(B,be)
Tense(B,%pres_ongoing)
objTARGET2(B,F)
subjDESCRIPTEE(B,B1)
Inheritance(B1,Amir)
Inheritance(F,friend)
ofDESCRIPTEE(F,O)
Inheritance(O,James)
```

Amir and James are friends

```
Inheritance(B,be)
Tense(B,%pres_ongoing)
objTARGET2(B,F)
subjDESCRIPTEE(B,group^1099074852934_3040)
```

```
Inheritance(B1,Amir)
Inheritance(B1,group^1099074852934_3040)
Inheritance(F,friend)
Inheritance(O,James)
Inheritance(O,group^1099074852934_3040)
```

Amir is James's friend

```
Inheritance(B,be)
Tense(B,%pres_ongoing)
objTARGET2(B,F)
subjDESCRIPTEE(B,O)
Inheritance(B1,Amir)
Inheritance(F,friend)
possFOCUS2(F,B1)
Inheritance(O,James)
```

These parses are produced by the rule-based INLINK system, but they are similar in form to the output one would expect to find from a learned NLP parsing schema within Novamente. However, there are two issues here:

1. The specific relationships, such as ofDESCRIPTEE and objTARGET, are too crude and broad. Finer-grained relationships along these same lines would be learned via experience in an embodied Novamente system, and would be more useful for guiding inference.
2. The first level of semantics derivable from English sentences is still too close to the English syntax, which means that differently-worded sentences may give rise to different-looking semantic representations. This sort of divergence of representation is problematic for inference: one would like Novamente be able to reason close to identically on knowledge derived from sentences that are close-to-identical in meaning.

The second problem may be handled by using a collection of semantic transformation rules to get from the language-ish knowledge representation exemplified above to a more inference- friendly representation. In general, at this stage, we require roughly one semantic transformation for each subject-argument relationship (e.g. subjAGENT) and each preposition sense (e.g. ofFOCUS) and also for senses of common "glue" verbs such as "be." These transformations are themselves represented as nodes and links and are executed via Novamente inference.

A simple example of such a transformation is the one for ofDESCRIPTEE, which looks like

```
ForAll R, X, Y: ImplicationLink( foo1, foo2)
foo2 = ( ofIze(R) )(X,Y)
foo1 = AND( ofDESCRIPTEE(R,Y), R(X) )
```

where *ofIze* is a Novamente SchemaNode corresponding to the meaning of the relevant sense of the word *of*.

The problem of the preposition and subject-argument-relation senses being too high-level and not refined enough is harder to solve. This is a problem that we believe can be most effectively addressed via having an AI system learn preposition and subject-argument relation and glue word senses based on experience. Certainly, statistical or rule-based approaches to linguistics have shown no particular capability for dealing with this sort of issue, so far. The INLINK approach appears to be the most sophisticated one in the literature, but it has known limitations. The Cyc knowledge base contains a fairly refined collection of meanings for these little words (e.g. 14 senses for "in), but it also comes nowhere near the subtlety and context-sensitivity needed for handling these words in a truly conceptually fluent way.

5.5 Human and Computational Perception

Human perceptual processing is a huge topic (Mountcastle, 1998) which we will not address in any depth here. We will merely give some brief indications of how we feel it can be handled within the conceptual and software frameworks outlined above.

In the human brain, each sense is handled somewhat differently. Different neural architectures are involved (Lynch, 1986). For instance the olfactory cortex has mainly combinatory connections, without any hierarchical structure, and is reasonably well understood via attractor neural network ideas (Freeman, 1997). On the other hand vision processing has a well known and complex hierarchical structure, which is genetically tuned to reduce the otherwise massive complexity of the information processing task it handles. This same diversity of structure will be needed in any AI system that intends to handle perceptual data processing within a reasonable amount of computational resources. But just as all the different senses are dealt with in the human brain using the same basic physiological mechanisms, similarly in an AI system they can all be dealt with using the same fundamental knowledge representation tools and learning dynamics.

In Novamente, raw percepts are represented by special node types, e.g. PixelNode for a pixel on a camera eye. Complex percepts are then represented as predicates combining raw percepts, embodied in PredicateNodes. The creation of appropriate predicates representing the structure of sensory input is handled by perception-processing schema, that in Novamente terms are "concept creation" schema. Conceptually, all this is straightforward, but the kicker is processing speed. Learning these concept creation schema would take infeasibly long, given the amount of data involved. Thus, the need for specialized architectures as are used in the human brain.

Due to its combinatory nature, olfactory processing may not need any special handling within Novamente; just some parameter-tuning. Learning categories of smells seems like a pure numerical supervised-classification/clustering type problem. On the other hand, vision's hierarchical structure obviously affects the inferences that one wants to bother doing when analyzing image data-- so to do vision processing effectively within Novamente, one needs a visual-hierarchy-guided inference control strategy.

One approach to this problem is to break space down into a hierarchy of cells (a "multiresolution hierarchy") and associate a separate pool of pattern-recognition schema

with each cell. In this approach each schema may have unrestricted structure, but the network of schemata dealing with an overall visual scene has a hierarchical structure imposed by the multiresolution hierarchy.

And, it may also be valuable to introduce biases to the schema learning process itself, depending specifically on the sense modality. For instance, in learning a schema dealing with an area of the visual field, you may want to search preferentially for dependencies between internal schema sub-nodes depending on one sub-area and internal schema sub-nodes depending on nearby sub-areas. And one can imagine subtler things along these lines being introduced, e.g. biasing the search to find dependencies between aspects of color, brightness, motion, etc. that are commonly interdependent.

All this gets quite technical – but the human psychology results show that it may be of importance beyond the domain of sense processing – especially the vision case, because in many contexts humans use visual imagery for cognitive purposes (Helstrop and Logie, 1999). Imagery may be used for retrieving subtle spatial or perceptual information from memory that hasn't been stored as such, and can't easily be deduced from other information; it may also be used for planning movements, understanding descriptions, and helping solve some kinds of problems. Of course, this is not a necessary prerequisite for advanced cognition – a non-visual AGI could still vastly exceed human intelligence – but it may be a prerequisite for humanlike cognition.

6 Novamente vs. the Human Brain/Mind: Self, Learning and Feeling

We have discussed a number of aspects of human psychology but we have left out a number of major areas – the core topics of "clinical psychology": self, awareness, will, emotion. These are slipperier topics than the more cognitive issues discussed above, but no less critical to human or AI intelligence. Now we will run through these various issues at a moderate level of depth, explaining how each one can be handled effectively within a SMEPH/Novamente framework, in a way that does justice to the best theories of human neuropsychology and also to the particular nature of digital computer systems.

6.1 Self and Intersubjectivity

According to the psynet model, a self is nothing mystical, it is a certain type of structure, evolving according to a certain type of dynamic, and depending on other structures and dynamics in specific ways. Self, we believe, is necessary for creative adaptability -- for the spontaneous generation of new routines to deal with new situations. Current AI programs do not have selves, and in fact they do not even have the component structures out of which selves are built. This is one of the reasons they aren't very intelligent.

We have argued (Goertzel, 1997) that the construction of the self is a key aspect of intelligence – and that the surest way to endow an AI with a self is to place is in a situation of "artificial intersubjectivity," where it gets to modify a (possibly simulated) world collaboratively with other intelligent agents.

6.1.1 The Nature of Self

What is the self? Psychology provides this question with not one but many answers. One of the most AI-relevant answers is given by Epstein's (1984) synthetic personality theory. Epstein argues that the self is a theory. This is a useful perspective for AI because theorization is something relatively well-understood within AI. This perspective fits in naturally with Metzinger's neurophilosophy-inspired notion of a "phenomenal self," mentioned above.

Epstein's personality theory paints a refreshingly simple picture of the mind:

> [T]he human mind is so constituted that it tends to organize experience into conceptual systems. Human brains make connections between events, and, having made connections, they connect the connections, and so on, until they have developed an organized system of higher- and lower-order constructs that is both differentiated and integrated. ...
> In addition to making connections between events, human brains have centers of pleasure and pain. The entire history of research on learning indicates that human and other higher-order animals are motivated to behave in a manner that brings pleasure and avoids pain. The human being thus has an interesting task cut out simply because of his or her biological structure: it is to construct a conceptual system in such a manner as to account for reality in a way that will produce the most favorable pleasure/pain ratio over the foreseeable future. This is obviously no simple matter, for the pursuit of pleasure and the acceptance of reality not infrequently appear to be at cross-purposes to each other.

He divides the human conceptual system into three categories: a self-theory, reality-theory, and connections between self-theory and reality-theory. And he notes that these theories may be judged by the same standards as theories in any other domain:

> [Since] all individuals require theories in order to structure their experiences and to direct their lives, it follows that the adequacy of their adjustment can be determined by the adequacy of their theories. Like a theory in science, a personal theory of reality can be evaluated by the following attributes: extensivity [breadth or range], parsimony, empirical validity, internal consistency, testability and usefulness.

A person's self-theory consists of her best guesses about what kind of entity she is. In large part it consists of ideas about the relationship between herself and other things, or herself and other people. Some of these ideas may be wrong; but this is not the point. The point is that the theory as a whole must have the same qualities required of scientific theories. It must be able to explain familiar situations. It must be able to

generate new explanations for unfamiliar situations. Its explanations must be detailed, sufficiently detailed to provide practical guidance for action. Insofar as possible, it should be concise and self-consistent.

The acquisition of a self-theory, in the development of the human mind, is intimately tied up with the body and the social network. The infant must learn to distinguish their body from the remainder of the world. By systematically using the sense of touch -- a sense that has never been reliably simulated in an AI program -- she grows to understand the relation between herself and other things. Next, by watching other people she learns about people; inferring that she herself is a person, she learns about herself. She learns to guess what others are thinking about her, and then incorporates these opinions into her self-theory. Most crucially, a large part of a person's self-theory is also a *meta-self-theory*: a theory about how to acquire information for one's self-theory. For instance, an insecure person learns to adjust her self-theory by incorporating only negative information. A person continually thrust into novel situations learns to revise her self-theory rapidly and extensively based on the changing opinions of others -- or else, perhaps, learns not to revise her self-theory based on the fickle evaluations of society.

We believe that capacity for creative intelligence is dependent on the possession of effective self- and reality- theories – because self- and reality- theories provide the *dynamic data structures* – the SMEPH derived-hypergraph subgraphs -- needed for flexible, adaptable, creative thought.

The single quality most lacking in current AI programs is the ability to go into a new situation and "get oriented." This is what is sometimes called the brittleness problem. Our AI programs, however intelligent in their specialized domains, do not know how to construct the representations that would allow them to apply their acumen to new situations. This general knack for "getting oriented" is something which humans acquire at a very early age.

People do not learn to get oriented all at once. They start out, as small children, by learning to orient themselves in relatively simple situations. By the time they build up to complicated social situations and abstract intellectual problems they have a good amount of experience behind them. Coming into a new situation, they are able to reason associatively: "What similar situations have I seen before?" And they are able to reason hierarchically: "What simpler situations is this one built out of?" By thus using the information gained from orienting themselves to previous situations, they are able to make reasonable guesses regarding the appropriate conceptual representations for the new situation. In other words, they build up a dynamic data structure consisting of new situations and the appropriate conceptual representations. This data structure is continually revised as new information that comes in, and it is used as a basis for acquiring new information. This data structure contains information about specific situation and also, more abstractly, about how to get oriented to new situations.

Now, we suspect that it is just not computationally feasible to learn how to get oriented to complex situations, without first having learned how to get oriented to simpler situations. This regress only bottoms out with the very simplest situations, the ones confronted by every human being by virtue of having a body and interacting with other humans. There is a natural order of learning here, which is, due to various psychological

and social factors, automatically followed by the normal human child. This natural order of learning is reflected, in the mind, by n hierarchical data structure in which more and more complex situations are comprehended in terms of simpler ones. But those who write AI programs have made little or no attempt to respect this natural order.

Typically, we provide our AI programs with concepts that "make no sense" to them, which they are intended to consider as given, a priori entities. On the other hand, to a human being, there are no given, a priori entities; everything bottoms out with the phenomenological and perceptual, with those very factors that play a central role in the initial formation of self- and reality-theories. To us, complex concepts and situations are made of simpler, related concepts and situations to which we already know how to orient ourselves; and this reduction continues down to the lowest level of sensations and feelings. To our AI programs, the hierarchy bottoms out prematurely, and thus there can be no functioning dynamic data structure for getting oriented, no creative adaptability, no true intelligence.

The way to get around these problems is to create AI programs that are not only embodied but also intrinsically social. This gives rise to the notion of artificial intersubjectivity or A-IS (Goertzel, 1997). The idea of A-IS is to simulate a *system of intelligences collectively creating their own subjective (simulated) reality.*

6.1.2 Artificial Intersubjectivity

The basic concept of A-IS is that a collection of artificially intelligent agents, in order to achieve a high level of intelligence via interacting in a simulated world, must collude in the modification of that world, so as to produce a mutually more useful simulated reality. In this way they may evolve interrelated self- and reality-theories, and thus artificial intersubjectivity.

The key question is whether this can be expected to happen spontaneously or not. This ties in with the human-psychology question of how much in-built mechanism we have for social modeling. While the jury is still out on the details, the correct answer seems to be “quite a lot” (Calvin and Bickerton, 2001).

So, it would seem that, speaking practically, spontaneously and automatic intersubjectivity cannot be counted on. Unless the different interacting AI agents are in some sense "wired for cooperativity," they may well never see the value of collaborative subjective-world-creation. We humans became intelligent in the context of collaborative world-creation, of intersubjectivity (even apes are intensely intersubjective). Unless one is dealing with AI agents that evolved their intelligence in a social context -- a theoretically possible but pragmatically tricky solution -- there is no reason to expect significant intersubjectivity to spontaneously emerge through interaction.

Fortunately, there is an alternative, which is the design strategy called "explicit socialization," which involves explicitly programming each AI agent in a community, from the start, with:

1) an a priori knowledge of the existence and autonomy of the other programs in its environment, and

2) an a priori inclination to model the behavior of these other programs.

In other words, in this strategy, one enforces A-IS from the outside, rather than, as in natural "implicit socialization," letting it evolve by itself. This approach is, to a certain extent, philosophically disappointing; but this may be the kind of sacrifice one must make in order to bridge the gap between theory and practice. In a Novamente context, what this boils down to is creating special Lobes that are explicitly purposed to serve as models for other minds: architecturally quite simple once one decides to do it. This aspect of Novamente is very well-suited for experimentation within the AGI-SIM simulation world.

6.2 Free Will

The problem of "free will" is a large one and conceals many subtleties and complexities (Dennett, 2003). However, recent neuroscience research by Benjamin Libet (2000), Michael Gazzaniga (1989) and others has shed substantial light on the matter from a practical perspective. Given these results, it is now possible to analyze in some depth how a phenomenon like "free will" may be realized in an AI system. In Metzinger's terms, we consider free will as an aspect of an intelligent system's constructed phenomenal self. Libet and Gazzaniga tell us some concrete and interesting things about the particularities of this aspect of the phenomenal self.

Free will, we suggest, has a lot to do with planning. The world is complex and uncertain, and an intelligent system rarely knows what's going to happen in the actual universe. So, in order to plan for the future, it must create a *virtual multiverse* inside itself: i.e. at time t it must model several different future states for time t+s, since it doesn't know which future state will actually occur. It must create a virtual multiverse with branch-points regarding its own external actions, and its own internal events, as well as external events not directly caused by itself. This is what our brains do all the time – and it is a process that arguably gives rise to a free-will-like experience within the phenomenal self.

We know that the cognitive portions of brains do not directly experience the external universe; they only experience their own models of the external universe. This is demonstrated by many experiments regarding perceptual illusions, for example (Maturana and Varela, 1992). What this means is that, even if we should happen to live in a strictly deterministic universe[7][1], we subjectively live in a multiverse in which several different possible branches are subjectively real at any given time. But most of these branches are very short-lived: they exist only conjecturally while we wait for the next percepts which will tell us which of the branches is actualized.

Furthermore, brains largely experience themselves only via their models of themselves. Brains, being complex systems, are hard to predict even for themselves, and so one part of a brain often must use a virtual multiverse to model another part.

When a brain triggers a real-world action, this action occurs in the external universe, and then registers internally in the virtual multiverse which models the external

universe. The brain is then aware of a process of "collapse" wherein the multiple branches of the virtual multiverse collapse to a single branch. Furthermore, this collapsing process occurs rapidly, within the same *subjectively experienced moment* as the actual event in the physical universe. Note that a subjectively experienced moment is not instantaneous.

Similarly, when a part of a brain carries out an action, and another part of the intelligent system is modeling this first part using a virtual multiverse, then the action in the first part corresponds with a collapse to a single branch in the virtual multiverse contained in the second part.

The special feeling of "free will" that we experience consists primarily of the subjectively-simultaneous consciousness of

- an event occurring in the external universe
- a collapse-to-a-single-branch occurring in the brain's internal virtual multiverse

or else the simultaneous consciousness of

- an event occurring in one part of the brain
- a collapse-to-a-single-branch occurring in the virtual multiverse used by another part of the brain to model the first part

The subjective simultaneity is only present when the two things occur at almost the same physical time, which generally occurs only when the event in question is either internal, or else an external event that's directly triggered by the brain itself.

Libet (2000) has done experiments showing that, in many cases, the "decision" to carry out an action occurs *after* the neural signals directly triggering the action have already occurred. This observation fits in perfectly with the virtual multiverse theory. Note that this time interval is sufficiently short that the action and the decision occur within the same subjectively experienced moment. In fact, Libet's results, though often presented as counterintuitive, are explained naturally by the current theory – it's the *opposite* result, that perceived-virtual-multiverse-collapses occurred *after* the corresponding actions, that would be more problematic for the current theory.

Dennett (2003) analyzes Libet's results by positing that free will is a distributed experience which occurs over an expanse of time (the experienced moment) and a number of different brain systems, and that there is nothing paradoxical about the part of this experience labeled "decision" occurring minutely before the part of this experience labeled "action trigger." I agree with Dennett's general observations – and with most of his comments about free will – but I am aiming to achieve a greater level of precision in my analysis of the phenomenon.

For example, suppose I am trying to decide whether to kiss my beautiful neighbor. One part of my brain is involved in a dynamic which will actually determine whether I kiss her or not. Another part of my brain is modeling that first part, and doesn't know what's going to happen. A virtual multiverse occurs in this second part of the brain, one branch in which I kiss her, the other in which I don't. Finally, the first part

comes to a conclusion; and the second part collapses its virtual multiverse model almost instantly thereafter.

The brain uses these virtual multiverse models to plan for multiple contingencies, so that it is prepared in advance, no matter what may happen. In the case that one part of the brain is modeling another part of the brain, sometimes the model produced by the second part may affect the actions taken by the first part. For instance, the part (call it B) modeling the action of kissing my neighbor may come to the conclusion that the branch in which I carry out the action is a bad one. This may affect the part (call it A) actually determining whether to carry out the kiss, causing the kiss not to occur. The dynamic in A which causes the kiss not to occur, is then reflected in B as a collapse in its virtual multiverse model of A.

Now, suppose that the timing of these two causal effects (from B to A and from A to B) is different. Suppose that the effect of B on A (of the model on the action) takes a while to happen (spanning several subjective moments), whereas the effect of A and B (of the action on the model) is nearly instantaneous (occurring within a single subjective moment). Then, another part of the brain, C, may record the fact that *a collapse to definiteness in B's virtual multiverse model of A, preceded an action in A.* On the other hand, the other direction of causality, in which the action in A caused a collapse in B's model of A, may be so fast that no other part of the brain notices that this was anything but simultaneous. In this case, various parts of the brain may gather the mistaken impression that virtual multiverse collapse causes actions; when in fact it's the other way around. This, I conjecture, is the origin of our mistaken impression that we make "decisions" that cause our actions.

The "illusion" of free will, therefore, consists largely of a mistaken impression gathered by some parts of the brain about the ordering of events in other parts of the brain. It consists of a confusion between two different roles played by virtual multiverse models:

- assisting in the determination of actions (which happens sometimes, and with a significant time lag)
- registering already-occurred actions (which happens more often, and almost instantaneously)

Because in the former, multiple-subjective-moment case, virtual multiverse collapse precedes action-determination, the brain mistakenly infers that in the latter, single-subjective-moment case, virtual multiverse collapse also precedes action-determination. But in fact, in the latter case virtual multiverse collapse follows action-determination.

However, it is not an illusion or confusion that virtual multiverse modeling has an impact on actions taken in the brain. This kind of modeling is clearly a very valuable part of brain dynamics, due to the complex and hard-to-predict nature of the brain and world. Virtual multiverse modeling is necessary due to *practical indeterminism* within and outside the brain, which exists whether or not *fundamental indeterminism* does. It is necessary because internal and external events are often *indeterministic from the subjective perspective of particular, useful parts of the brain.* Furthermore, and critically, *the brain as a whole is often indeterministic from its own perspective.*

Finally, the phenomenon of confabulation (Gazzaniga, 1989) adds a third aspect to virtual multiverse dynamics: not only do virtual multiverse inferences/simulations affect actions, and actions cause updating of virtual multiverse simulations; but also, reasoning about actions causes inferred stories to be attached to the memories of virtual-multiverse collapses.

6.3 Awareness

Free will leads us naturally to the even thornier issue of "consciousness" or "awareness" – an issue more controversial by far than anything else in the cognitive science domain. The SMEPH/Novamente approach is neutral as regards the ultimate nature of consciousness, although the author has his own opinions (Goertzel, 2004d). However, the approach to free will described above partially addresses the phenomenon of consciousness, in a way that we'll briefly outline here.

Some aspects of consciousness can be understood by thinking about the virtual multiverse models that parts of the brain construct, in order to model the brain as a whole. These virtual multiverse models are used to help guide the dynamics of the whole brain (on a slow time scale), and they are also continually updated to reflect the actual dynamics of the brain (on a faster time scale, occurring within a single subjective moment). The feeling of consciousness is in part the feeling of events in the whole brain being rapidly reflected in the changes in the virtual multiverse models maintained in parts of the brain … and these changes then causing further virtual-multiverse-model changes which then feed back to change the state of the whole brain again … etc. The conscious feeling of the flow of time is actually a feeling of continual ongoing branch-selection in the virtual multiverse model of the whole brain – the feeling of briefly-explored possible futures being left by the wayside as the actualized futures are registered in the model.

Dennett (1992) analyzed human consciousness as a serial computer running as a virtual machine on top of a parallel computer (the "parallel computer" being the unconscious, which comprises the majority of brain function). However, I don't think this is quite right. Rather, I think human consciousness has to do with the feedback between virtual multiverse modeler software (embodied in various parts of the brain) and massively parallel software (the rest of the brain). The virtual universe modeler software is not exactly a serial computation process, it may well explore multiple branches in parallel.

The virtual-multiverse theory of free will does not explicitly solve the "hard problem of consciousness" (Chalmers, 1997), the relationship between subjective awareness ("qualia") and physical phenomena. However, it does fit in naturally with a particular hypothetical solution to the hard problem. Suppose one accepts, as a solution to the hard problem, the postulate that *a quale occurs when a system comes to display a pattern that it did not display a moment before; and the more prominent patterns correspond to the more intense qualia.* Then, it follows from the present theory of free will that intense qualia will tend to be correlated with significant activity in the whole-brain virtual multiverse modeler. This provides an explanation for the oft-perceived

correlation between consciousness and free will (free will also often being associated with significant activity in the whole-brain virtual multiverse modeler).

6.4 Emotion

Following "self" and "awareness," another critical aspect of human "clinical psychology" is emotion. Emotions play an extremely important role in human mental life – but it is not, on the face of it, clear whether this needs to be the case for AI's. Much of human emotional life is distinctly *human* in nature, clearly not portable to systems without humanlike bodies. Furthermore, many problems in human psychology and society are caused by emotions run amok in various ways – so in respects it might seem desirable to create emotion-free AI's.

On the other hand, it may also be that emotions represent a critical part of mental process, and human emotions are merely one particular manifestation of a more general phenomenon – which must be manifested in *some* way in any mind. This is the perspective we'll advocate here. We suggest that the basic phenomenon of emotion is something that any mind must experience. Human emotions are then considered as an elaboration of the general "emotion" phenomenon in a peculiarly human way. There are a few universal emotions – including happiness, sadness and spiritual joy – which any intelligent system with finite computational resources is bound to experience, to an extent. And then there are many species-specific emotions, which in the case of humans include rage, joy and lust and other related feelings.

Generally speaking motions have two aspects, which may be called *hot* versus *cold* (Mandler, 1975), or "conscious-experiential-flavor" versus "neural/cognitive structure-and-dynamics" – or, using our preferred vocabulary, *qualia* versus *pattern*. From some conceptual perspectives, the relation between the qualia aspects and the pattern aspect is problematic. We prefer a philosophy in which qualia and patterns are aligned – each pattern comes along with a quale, which is more or less intense according to the "prominence" of the pattern (the degree of simplification that the pattern provides in its ground) (see Goertzel, 2004a). In this approach, the qualia and pattern aspects of emotion may be dealt with in a unified way.

So what is the general pattern of "emotion"? The working conceptual definition given in (Goertzel, 2004c) is as follows:

A mental state that does not arise through free will, and that and is often accompanied by physiological changes

"Free will," as proposed in (Goertzel, 2004b), is a complex sort of quale, consisting primarily of

- the registration of an (internal or external) action in an intelligent system's "virtual multiverse model," roughly simultaneously with the execution of that action

and generally going along with

- the construction of causal models explaining what internal structures and dynamics caused the action

Sometimes these two aspects are uncorrelated, giving the feeling of "I don't know why I decided to do that."

Mental states that do not arise through free will, are mental states that are registered in the virtual multiverse model only considerably after they have occurred, thus giving a feeling of "having spontaneously arisen". This often goes along with arising through such a large-scale and complicated – or opaque -- process that detailed causal modeling is difficult. But sometimes, these two aspects are uncorrelated, and one can rationally reconstruct why some spontaneous mind-state occurred, in a reasonably confident way.

What causes mental states to register in the brain's virtual multiverse model in a delayed way? One cause might be that these mental states are ambiguous and difficult to understand, so that it takes the virtual multiverse modeler a long time to understand what's going on – to figure out which branch has actually been traversed. Another might be that the state is correlated with physical processes that inhibit the virtual multiverse modeler's normal "branch collapsing" activity – and that the branch-collapsing only proceeds a little later, once this inhibitory effect has diminished.

In the case of human emotions, the "accompaniment with physiological changes" mentioned in the above definition of emotion seems to be a key point. It seems that there's a time lag between *certain kinds of broadly-based physiological sensations in the human brain/body,* and *registration of these sensations in the human brain's virtual multiverse modelers.*

And so, in regard to emotions, a flexibly superposed subjective multiverse is maintained, rather than a continually collapsed subjective universe that defines a single crisp path through the virtual multiverse. This helps explain both the beautiful and the confusing nature of emotions.

Regarding the second hypothesized factor, the obvious question is: Why do the broadly-based partly-physical sensations we humans call "emotions" have this strange relationship with time? This may be largely because they consist of various types of data coming in from various parts of the brain and body, with various time lags. A piece of sensation coming in from one part of the brain or body right now may have a different meaning depending on information about what's going on in some other part of the brain or body – but this information may not be there yet. When information gathering and integration regarding a "distributed action pattern" requires this kind of temporally-defused activity, then the tight connection between action and virtual-multiverse-model collapse that exists in other contexts doesn't exist anymore. Ergo, no feeling of "free will" – rather, a feeling of things happening in oneself, without a correlated "decision process." A strong emotion can make one feel "outside of time."

Furthermore, while it's easy to make a high-level story as to what made one sad or happy or feel some other emotion, it's not at all easy to make up a story regarding the details of an emotional experience. Usually, one just doesn't know – because so much of the details of the emotional experience have to do with physiological dynamics that are opaque to the analytical brain (unless the analytical brain makes a huge, massively-effort-consuming push to become aware of these normally unconscious processes).

These comments lead us to a more specific, technical, "mechanistic" and hypothetical definition of emotion:

A mental state marked by prominent internal temporal patterns that

- *are not controllable to any reasonable extent by the virtual multiverse modeling subsystem, or*
- *have the property that their state at each time is far more easily interpretable by integration of past and future information.*

Such patterns will often, though not always, involve complex and broad physiological changes.

And what does this mean regarding the potential experiencing of emotions by nonhuman minds? Clearly, in any case where there's diverse and ambiguous information coming in from various hard-to-control parts of an intelligent system, one is not going to have the "usual" situation of virtual multiverse collapse. One is going to have a sensation of major patterns occurring inside one's own mind, but without any "free will" type "decision" process going along with it. This is, in the most abstract sense, "emotion." Emotions in this sense need not be correlated with physiological patterns, but it makes sense that they often will be.

This also brings up the question of emotional typology. Humans experience a vast range of emotions. Will other types of minds experience completely different emotion-types, or is there some kind of general system-theoretic typology of emotions?

The line of thinking pursued here suggests that there will be a small amount of emotional commonality among various minds – certain very simple emotions have an abstract, mind-architecture-independent meaning. But the vast majority of human emotional nuance is tied to human physical embodiment and evolutionary history, and would not be emulated in an AI mind or a radically different biological species.

For instance, any system that has a set of goals that remain constant over a period of time, can experience an emotion we may call "abstract happiness," which is ***the emotion induced by an increasing amount of goal-achievement***. On the other hand, it can also experience "abstract sadness," i.e. ***the emotion induced by a decreasing amount of goal-achievement***. These emotions can become quite complex because organisms can have multiple goals, and at any one moment some may experience increasing achievement while others experience decreasing achievement.

What of specific human emotions like lust, rage and fear? Clearly these exist because we have specific physiological response systems for dealing with specific situations. Fear activates flight-related subsystems; rage activates battle-related subsystems; lust activates sex-related subsystems. Each of these body subsystems, when activated, floods the brain with intensive and diverse and hard-to-process stimuli, which

are beyond the control of "free will" related processes. Many of the responses of these body subsystems are fast -- too fast for virtual multiverse modeling to deal with. They're fast because primordially they had to be fast – you can't always stop to ponder before running, attacking or mating.

All in all, there's no doubt that, unless an AI system is given a mammal-like motivational system, its emotional makeup will vastly differ from that of humans. An AI system won't necessarily have strong emotions associated with battle, reproduction or flight. Conceivably it could have subsystems associated with these types of actions, but even so, it could be given a much greater ability to introspect into these subsystems than humans have in regard to their analogous subsystems.

So the most reasonable conclusion about AI emotions is that:

- AI systems clearly will have emotions
- Their emotions will include, at least, happiness and sadness and spiritual joy
- Generally AI systems will probably experience less intense emotions than humans, because they can have more robust virtual multiverse modeling components, which are not so easily bollixed up – so they'll less often have the experience of major non-free-will-related mental-state shifts
- Experiencing less intense emotions does not imply experiencing less intense states of consciousness. Emotion is only one particular species of state-of-consciousness.
- The specific emotions AI systems will experience will probably be quite different from those of humans, and will quite possibly vary widely among different AI systems
- If you put an AI in a human-like body with the same sorts of needs as primordial humans, it would probably develop every similar emotions to the human ones

We now briefly consider these issues in terms of the specific structures and dynamics of the Novamente AI system. In this context, a specific prediction made by the present theory of emotions is that complex map dynamics will be more associated with emotions than other aspects of Novamente cognition. Complex map dynamics involve temporal patterns that are hard to control, and that present sufficiently subtle patterns that the present is much better understood once one knows the immediate future. One may infer from this a possible major feature of the difference between Novamente psychology and human psychology: the strongest emotions of a Novamente system may be associated with the most complexly unpredictable cognitions it has -- rather than, in humans, with phenomena that evoke the activities of powerful, primordial, opaque-to-cognition subsystems.

References

- Adams, Bryan, Cynthia Breazeal, Rodney Brooks, and Brian Scassellati. Humanoid Robots: A New Kind of Tool, *IEEE Intelligent Systems*, Vol. 15, No. 4, July/August, 2000, pp. 25—31.
- Alexander, R. McNeill (1996). Optima for Animals. Princeton University Press.
- Amit, Daniel (1999). *Modeling Brain Function*, Cambridge University Press
- Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human Computer Interaction*, 12(4), 439-462.
- Anderson, J. R. *Learning and Memory*. 2nd ed. N.Y.: Wiley, 2000. Chap. 2 and some of 3
- Baddeley, Alan (1999). *Essentials of Human Memory*. Taylor and Francis Group.
- Baum, Eric (2004). *What Is Thought?*, Bradford
- Berge, Claude (1999). *Hypergraphs*, Elsevier Science
- Boroditsky, L. & Ramscar, M. (2002). The Roles of Body and Mind in Abstract Thought. Psychological Science, 13(2), 185-188
- Cabeza, Roberto and Alan Kingstone (2001). *Handbook of Functional Neuroimaging of Cognition*, The MIT Press
- Calvin, William and Derek Bickerton (2001). *Lingua ex Machina*. MIT Press.
- Chalmers, David (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3):200-19, 1995
- Curry, Haskell and Robert Feys (1958). *Combinatory Logic*, North-Holland
- de Garis, Hugo, and Michael Korkin (2002). THE CAM-BRAIN MACHINE (CBM): An FPGA Based Hardware Tool which Evolves a 1000 Neuron Net Circuit Module in Seconds and Updates a 75 Million Neuron Artificial Brain for Real Time Robot Control, *Neurocomputing*, Elsevier, Vol. 42, Issue 1-4,
- Dennett, Daniel (1992). *Consciousness Explained.* Viking.
- Dennett, Daniel (2003). *Freedom Evolves*. Viking.
- Douthwaite, Julia V. (1997). *The Wild Girl, Natural Man and the Monster: Dangerous Experiments in the Age of Enlightenment*, University of Chicago Press
- Edelman, Gerald (1987). *Neural Darwinism*, Basic Books
- Epstein, Seymour (1994). *Integration of the Cognitive and the Psychodynamic Unconscious*. American Psychologist. Vol. 49, No. 8, 709-724.
- Freeman, Walter (2001). *How Brains Make Up Their Minds.* Columbia University Press.
- Gazzaniga, Michael (1989). "Organization of the Human Brain," *Science*, Sept., pp. 947-956
- Gleason, Jean (2004). *The Development of Language*. Allyn and Bacon.
- Goertzel, Ben (1993). *The Structure of Intelligence*, Springer-Verlag
- Goertzel, Ben (1993a). *The Evolving Mind*, Gordon & Breach
- Goertzel, Ben (1994). *Chaotic Logic*, Plenum
- Goertzel, Ben (1997). *From Complexity to Creativity*, Plenum

- Goertzel, Ben (2001). *Creating Internet Intelligence*, Plenum
- Goertzel, Ben et al (2003). Novamente: An Integrative Architecture for Artificial General Intelligence. *Proceedings of IJCAI-03 Workshop on Agents and Cognitive Modeling*, Acapulco, August 2003
- Goertzel, Ben (2003a). Mindplexes. *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2003/mindplex.htm
- Goertzel, Ben (2004a). Patterns of Awareness. *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2004/HardProblem.htm
- Goertzel, Ben (2004b). The Virtual Multiverse Theory of Free Will.. *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2004/FreeWill.htm
- Goertzel, Ben (2003). Hebbian Logic Networks, *Dynamical Psychology*, www.**goertzel**.org/dynapsyc/2003/HebbianLogic03.htm
- Goertzel, Ben, Moshe Looks and Cassio Pennachin (2004). Novamente: An Integrative Architecture for Artificial General Intelligence. *Proceedings of AAAI Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research*, Washington DC, August 2004
- Goertzel, Ben (in prep.). The Hidden Pattern
- Goertzel, Ben and Cassio Pennachin (in prep.). Foundations of Emergent Cognition: Modeling and Engineering General Intelligence Using Self-Modifying, Evolving Probabilistic Hypergraphs
- Goertzel, Ben, Matthew Ikle' and Izabela Goertzel (in prep.). Probabilistic Term Logic.
- Goertzel, Ben, Moshe Looks, Michael Ross, Cassio Pennachin and Hugo Pinto (2005). *NL Comprehension via Integrative AI and Human-Computer Interaction*, submitted for publication
- Goertzel, Ben, Cate Hartley, Ken Silverman, Michael Ross and Stephan Bugaj (2002). The Baby Webmind Project, Proceedings of AISB 2000
- Goldstone, R. L., Feng, Y., & Rogosky, B. (2005). In D. Pecher & R. Zwaan (Eds.) *Grounding cognition: The role of perception and action in memory, language, and thinking.* Cambridge: Cambridge University Press.
- Grossberg, Stephen (1987). *The adaptive brain, I and II***,** Amsterdam: North-Holland, 1987.
- Haikkonen, Pentti (2003). The Cognitive Approach to Conscious Machines. Imprint Academics
- Hawkins, Jeff (2004). *On Intelligence.* Times Books.
- Helstrop, Tore and Robert Logie (1999). *Imagery in Working Memory and Mental Discovery*, Taylor and Francis Group.
- Holland, John (1992). *Adaptation in Natural and Artificial Systems*, MIT Press
- Jones, R. Tambe, M., Laird, J., Rosenbloom, P. (1993). Intelligent automated agents for flight training simulators. In *Proceedings of the Third Conference on Computer Generated Forces and Behavioral Representation*. University of Central Florida. IST-TR-93-07.

- Hunt, R. R., & Ellis, H. C. (1999). Fundamentals of cognitive psychology - 6th Edition.
- Koza, John (1992). *Genetic Programming*. MIT Press.
- Laird, J.E., A. Newell, and P. S. Rosenbloom (1987). Soar: An architecture for general intelligence. Artificial Intelligence, 33(1):1--6. Luck, S. J., E. K. Vogel, and K. L. Shapiro, "Word Meanings Can be Accessed but not Reported During the Attentional Blink." *Nature*, 383 (October 17, 1996): 616-618.
- Lenat, Doug and R.V. Guha (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Projec*t, Addison-Wesley
- Libet, B., A. Freeman and K. Sutherland (2000). *The Volitional Brain: Towards a Neuroscience of Free Will*. Imprint Academic.
- Luck, S.J., Vogel, E. K. and Shapiro, K. L. (1996). *Word meanings can be accessed but not reported during the attentional blink.* Nature 383 616-618)
- Lynch, G. (1986). *Synapses, Circuits, and the Beginnings of Memory*, MIT Press, Cambridge, Massachusetts.
- Mandler, George (1975). *The Psychology of Emotion.* Wiley.
- Manning, Christopher and Heinrich Schutze (1999). *Foundations of Statistical Language Processing*. MIT Press.
- Maturana, Humberto and Francisco Varela (1992). *The Tree of Knowledge*. Shambhala.
- Metzinger, Thomas (2004). *Being No One*. Cambridge MA, MIT Press.
- Mountcastle, Vernon (1998). *Perceptual Neuroscience: The Cerebral Cortex*, Harvard University Press
- Nicholas, Nick and John Cowan (2003). *What Is Lojban?* , Logical Language Group
- Peirce, C. S. (1892) "The Law of Mind." Reprinted in Hartshorne, Charles, Paul Weiss and Arthur W. Burks eds, *The Collected Papers of Charles Sanders Peirce*, Cambridge: Harvard University Press, 1980, 6.102-6.163.
- Pelikan, Martin (2002). *The Bayesian Optimization Algorithm: From Single Level to Hierarchy*, PhD Thesis, CS Dept., UIUC
- Pennachin, Cassio, Lucio Coelho, Murilo Saraiva, Francisco Lobo, Francisco Prosdocimi, Kenji Shikida Ben Goertzel (2005). *Knowledge-Guided Analysis of Gene Expression Data Using Genetic Programming, Support Vector Machines and the Gene-Ontology and PIR Databases*, submitted for publication
- Piaget, Jean, Malcom Piercy and D.E. Berlin (2001). *The Psychology of Intelligence*. Routledge.
- Pietelli-Palmarini, Massamo (1996). *Inevitable Illusions*. Wiley.
- Reisberg, D. *Cognition: Exploring the Science of the Mind.* 2nd ed. New York: Norton, 2001.
- Reigler, A. (2005). Constructive Memory. *Kybernetes* vol.34, nos. 1/2, 2005, pp. 89-104.
- Robinson, Abraham and Andrei Voronkov (2001). *Handbook of Automated Reasoning*. MIT Press.

- Rowan, John (1990). *Subpersonalities: The People Inside Us,*. Routledge
- Roy, Deb and Niloy Mukherjee (2005). Towards Situated Speech Understanding: Visual Context Priming of Language Models. Computer Speech and Language, 19(2), pages 227-248
- Santore, John F. and Stuart C. Shapiro (2003). Crystal Cassie: Use of a 3-D Gaming Environment for a Cognitive Agent. In R. Sun, Ed., *Papers of the IJCAI 2003 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico, August 9, 2003, 84-91
- Shapiro, Stuart (2000). An Introduction to SNePS 3. In Bernhard Ganter & Guy W. Mineau, Eds. *Conceptual Structures: Logical, Linguistic, and Computational Issues.* Lecture Notes in Artificial Intelligence 1867. Springer-Verlag, Berlin, 2000, 510-524.
- Singer, W. (2001). Consciousness and the Binding Problem. *Ann N Y Acad Sci.* 2001 Apr;929:123-46.
- Solomonoff, Ray (1964). "A Formal Theory of Inductive Inference, Part I", Information and Control, Part I: Vol 7, No. 1, pp. 1-22, March 1964.
- Solomonoff, Ray (1964a). "A Formal Theory of Inductive Inference, Part II", Information and Control, Part II: Vol. 7, No. 2, pp. 224-254, June 1964a.
- Sommers, Frederic, George Englebretsen and Harry Wolfson (2000). An Invitation to Formal Reasoning. Ashgate Publishing.
- Sutton, Richard and Andrew S. Barto (1998). *Reinforcement Learning*. MIT Press.
- Underwood, G, ed. (1993). *The Psychology of Attention*. Vol 1. Aldershot: Elgar.
- Voss, Peter (2005). The Essentials of General Intelligence. In Goertzel and Pennachin (Ed.), *Artificial General Intelligence*, Springer-Verlag
- Wang, Pei (1995). *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence*. PhD Thesis, University of Indiana, 1995.
- Wason, P.C. (1966). Reasoning. In B. M. Foss (Ed.) *New Horizons in Psychology*, Penguin
- Wilson, Paul (2000). Refining the Neuro-Connector Model, MSc. Thesis, Division of Informatics, University of Edinburgh, http://www.inf.ed.ac.uk/publications/thesis/online/IM000112.pdf
- Wixted, John (2004). The Psychology and Neuroscience of Forgetting, Annual Review of Psychology, Vol. 55: 235-269