The CogPrime Architecture for Embodied Artificial General Intelligence

Ben Goertzel
Novamente LLC 1405 Bernerd Place
Rockville MD 20851, USA
Email: ben@goertzel.org

Shujing Ke[1] and Ruiting Lian[1,2]
and Jade O'Neill[1] and Keyvan Sadeghi [1]
and Dingjie Wang[1,2] and Oliver Watkins[1] and Gino Yu[1]
[1]School of Design
Hong Kong Poly U
[2] Department of Cognitive Science
Xiamen University, China

***Abstract*—CogPrime, a comprehensive architecture for embodied Artificial General Intelligence, is reviewed, covering the core architecture and algorithms, the underlying conceptual motivations, and the emergent structures, dynamics and functionalities expected to arise in a completely implemented CogPrime system once it has undergone appropriate experience and education. A qualitative argument is sketched, in favor of the assertion that a completed CogPrime system, given a modest amount of experience in an embodiment enabling it to experience a reasonably rich human-like world, will give rise to human-level general intelligence (with significant difference from humans, and with potential for progress beyond this level).**

I. Introduction

This paper reviews CogPrime, a conceptual and technical architecture for artificial general intelligence (AGI, [1]), founded on a systems theory of mind [2]. While the practical implementation and testing of CogPrime is still at an early stage, the goal of CogPrime is to manifest the same qualitative sort of general intelligence as human beings; and ultimately to be extendable to an even broader scope of intelligent functions.

To allay potential confusion we offer two caveats from the outset. First, CogPrime is not a model of human neural or cognitive structure or activity. It draws heavily on knowledge about human intelligence, especially cognitive psychology; but it also deviates from the known nature of human intelligence in many ways, with a goal of providing maximal humanly-meaningful general intelligence using available computer hardware. Second, CogPrime is not proposed as the one and only holy grail path to advanced AGI. We feel confident there are multiple possible paths to advanced AGI, and that in following any of these paths, multiple theoretical and practical lessons will be learned, leading to modifications of the ideas developed and possessed along the early stages of the path. The goal here is to articulate **one** path that we believe makes sense to follow, one overall design that we believe can work, for achieving general intelligence that is qualitatively human-level and in many respects human-like, without emulating human neural or cognitive function in detail.

CogPrime is described in more detail in a forthcoming book titled *Building Better Minds: The CogPrime Architecture for Artificial General Intelligence* [3], which exceeds 1000 pages including appendices; the goal of this paper is to outline some of the key points in a more compact format. For space reasons we omit here a detailed comparison of CogPrime with other existing cognitive architectures; the reader is referred to our papers [4], [5] which review the landscape of artificial brain and AGI architectures, and include a comparison to an earlier version of CogPrime.

A. Cognitive Synergy: A Central Design Concept in CogPrime

The last decades have seen dramatic developments in various technologies allied to, and supportive of, AGI, including computing hardware and software, computer science algorithms, cognitive science and neuroscience [6]. There is no consensus on why all this dramatic progress has not yet yielded AI software systems with anything near human-like general intelligence. However, we hypothesize that the core reason boils down to the following three points:

- Intelligence depends on the emergence of certain high-level structures and dynamics across a system's whole knowledge base;
- We have not discovered any one algorithm or approach capable of yielding the emergence of these structures;
- Achieving the emergence of these structures within a system formed by integrating a number of different AI algorithms and structures is tricky. It requires careful attention to the manner in which these algorithms and structures are integrated; and so far the integration has not been done in the correct way.

The human brain appears to be an integration of an assemblage of diverse structures and dynamics, built using common components and arranged according to a sensible cognitive architecture. However, its algorithms and structures have been honed by evolution to work closely together – they are very tightly inter-adapted, in somewhat the same way that the different organs of the body are adapted to work together. Due their close interoperation they give rise to the overall systemic behaviors that characterize human-like general intelligence. We believe that the main missing ingredient in AI so far is **cognitive synergy**: the fitting-together of different intelligent components into an appropriate cognitive architecture, in such a way that the components richly and dynamically support and assist each other, interrelating very closely in a similar manner

978-1-4673-5923-8/13/$31.00

to the components of the brain or body and thus giving rise to appropriate emergent structures and dynamics. Which leads us to one of the central hypotheses underlying the CogPrime approach to AGI: that **the cognitive synergy ensuing from integrating multiple symbolic and subsymbolic learning and memory components in an appropriate cognitive architecture and environment, can yield robust intelligence at the human level and ultimately beyond.**

The reason this sort of intimate integration has not yet been explored much is that it's difficult on multiple levels, requiring the design of an architecture and its component algorithms with a view toward the structures and dynamics that will arise in the system once it is coupled with an appropriate environment. Typically, the AI algorithms and structures corresponding to different cognitive functions have been developed based on divergent theoretical principles, by disparate communities of researchers, and have been tuned for effective performance on different tasks in different environments. Making such diverse components work together in a truly synergetic and cooperative way is a tall order, yet we believe that this – rather than some particular algorithm, structure or architectural principle – is the "secret sauce" needed to create human-level AGI based on technologies available today.

B. What Kind of "Intelligence" is CogPrime Aimed At?

We have mentioned "intelligence" frequently in the preceding paragraphs, but haven't specified precisely what we mean by it. In fact, a host of different definitions of intelligence have been proposed in the AGI, narrow AI, psychology, engineering and philosophy communities; Legg and Hutter [7] have enumerated over 70. We pragmatically conceive of general intelligence as "the ability to achieve complex goals in complex environments, using limited resources"; and we accept that any real-world general intelligence is going to be more intelligent with respect to some sorts of goals and environments than others. In [8] we have outlined a mathematical definition of intelligence that embodies this intuition. However, the CogPrime design is not bound to the particulars of this definition of intelligence, nor any other. CogPrime has been designed with careful attention toward human-relevant goals and environment, yet is also expected to have a particular set of intellectual strengths and weaknesses different from that of humans. CogPrime has been primarily motivated by a commonsensical interpretation of intelligence, comprising a mixture of theoretical general problem solving capability, and the practical ability to display the same sorts of intelligence as humans do.

II. CogPrime and OpenCog

The CogPrime architecture is closely allied with the OpenCog open-source AI software framework; however the two are not synonymous. OpenCog is a more general framework, suitable for implementation of a variety of specialized AI applications as well as, potentially, alternate AGI designs. And CogPrime could potentially be implemented other than within the OpenCog framework. The particular implementation of CogPrime in OpenCog is called OpenCogPrime. OpenCog was designed with the purpose, alongside others, of enabling efficient, scalable implementation of the full CogPrime design.

A. Current and Prior Applications of OpenCog

To give greater understanding regarding the practical platform for current work aimed at realizing CogPrime, we now briefly discuss some of the particulars of work done with the OpenCog system that currently implements parts of the CogPrime architecture.

OpenCog, the open-source software framework underlying the "OpenCogPrime" (currently partial) implementation of the CogPrime architecture, has been used for commercial applications in the area of natural language processing and data mining. For instance, see [2] where OpenCogPrime's PLN reasoning and RelEx language processing are combined to do automated biological hypothesis generation based on information gathered from PubMed abstracts. [9] describes the use of OpenCog's MOSES component for biological data analysis; this use has been extended considerably in a variety of unpublished commercial applications since that point, in domains such as financial prediction, genetics, marketing data analysis and natural language processing. Most relevantly to the present work, OpenCog has also been used to control virtual agents in virtual worlds [10].

Prototype work done during 2007-2008 involved using an OpenCog variant called the OpenPetBrain to control virtual dogs in a virtual world. While these OpenCog virtual dogs did not display intelligence closely comparable to that of real dogs (or human children), they did demonstrate a variety of interesting and relevant functionalities including

- learning new behaviors based on imitation and reinforcement
- responding to natural language commands and questions, with appropriate actions and natural language replies
- spontaneous exploration of their world, remembering their experiences and using them to bias future learning and linguistic interaction

One current OpenCog initiative [11] involves extending the virtual dog work via using OpenCog to control virtual agents in a game world inspired by the game Minecraft. These agents are initially specifically concerned with achieving goals in a game world via constructing structures with blocks and carrying out simple English communications. Representative example tasks would be:

- Learning to build steps or ladders to get desired objects that are high up
- Learning to build a shelter to protect itself from aggressors
- Learning to build structures resembling structures that its shown (even if the available materials are a bit different)
- Learning how to build bridges to cross chasms

Of course, the AI significance of learning tasks like this all depends on what kind of feedback the system is given,

and how complex its environment is. It would be relatively simple to make an AI system do things like this in a highly specialized way, but that is not the intent of the project – the goal is to have the system learn to carry out tasks like this using general learning mechanisms and a general cognitive architecture, based on embodied experience and only scant feedback from human teachers. If successful, this will provide an outstanding platform for ongoing AGI development, as well as a visually appealing and immediately meaningful demo for OpenCog.

A few of the specific tasks that are the focus of this project teams current work at time of writing include:

- Watch another character build steps to reach a high-up object
- Figure out via imitation of this that, in a different context, building steps to reach a high up object may be a good idea
- Also figure out that, if it wants a certain high-up object but there are no materials for building steps available, finding some other way to get elevated will be a good idea that may help it get the object (including e.g. building a ladder, or asking someone tall to pick it up, etc.)
- Figure out that, if the character wants to hide its valued object from a creature much larger than it, it should build a container with a small hole that the character can get through, but the creature cannot

B. Transitioning from Virtual Agents to a Physical Robot

In 2009-2010, preliminary experiments were conducted using OpenCog to control a Nao robot [12]. These involved hybridizing OpenCog with a separate subsystem handling low-level perception and action. This hybridization was accomplished in an extremely simplistic way, however. How to do this right is a topic treated in detail in [3] and [13] and only briefly touched here. Work in this direction will proceed during 2013 and 2014, funded by a grant from the Hong Kong Innovation in Technology Fund.

We suspect that reasonable level of capability will be achievable by simply interposing DeSTIN [14] (or some other reasonably capable ”hierarchical temporal memory” type sensorimotor system) as a perception/action “black box” between OpenCog and a robot. However, we also suspect that to achieve robustly intelligent robotics we must go beyond this approach, and connect robot perception and actuation software with OpenCogPrime in a “white box” manner that allows intimate dynamic feedback between perceptual, motoric, cognitive and linguistic functions. We suspect this may be achievable, for example, via the creation and real-time utilization of links between the nodes in CogPrime ’s and DeSTIN’s internal networks.

III. Conceptual Background

The creation of advanced AGI systems is an engineering endeavor, whose achievement will require significant input from science and mathematics; and also, we believe, guidance from philosophy. Having an appropriate philosophy of mind certainly is no guarantee of creating advanced AGI system; philosophy only goes so far. However, having a badly inappropriate philosophy of mind may be a huge barrier in the creation of AGI systems.

The development of the CogPrime design has been substantially guided by a philosophy of mind called ”patternism” [15]. The patternist philosophy of mind is a general approach to thinking about intelligent systems. It is based on the very simple premise that *mind is made of pattern* – and that a mind is a system for recognizing patterns in itself and the world, critically including patterns regarding which procedures are likely to lead to the achievement of which goals in which contexts. However, the guidance provided to CogPrime by the patternist perspective should not be overstated. CogPrime is an integrative design formed via the combination of a number of different philosophical, scientific and engineering ideas, and the success or failure of the design doesn’t depend on any particular philosophical understanding of intelligence.

Pursuing the patternist philosophy in detail leads to a variety of particular hypotheses and conclusions about the nature of mind. Following from the view of intelligence in terms of achieving complex goals in complex environments, comes a view in which the dynamics of a cognitive system are understood to be governed by two main forces:

- self-organization, via which system dynamics cause existing system patterns to give rise to new ones
- goal-oriented behavior, which has been defined more rigorously in [8], but basically amounts to a system interacting with its environment in a way that appears like an attempt to maximize some reasonably simple function

Self-organized and goal-oriented behavior must be understood as cooperative aspects. For instance – to introduce an example that will be elaborated in more detail below – an agent is asked to build a surprising structure out of blocks and does so, this is goal-oriented. But the agent’s ability to carry out this goal-oriented task will be greater if it has previously played around with blocks a lot in an unstructured, spontaneous way. And the “nudge toward creativity” given to it by asking it to build a surprising blocks structure may cause it to explore some novel patterns, which then feed into its future unstructured blocks play.

Based on these concepts, as argued in detail in [15], several primary dynamical principles may be posited, including the following.

- **Evolution** , conceived as a general process via which patterns within a large population thereof are differentially selected and used as the basis for formation of new patterns, based on some “fitness function” that is generally tied to the goals of the agent
- **Autopoiesis**: the process by which a system of interrelated patterns maintains its integrity, via a dynamic in which whenever one of the patterns in the system begins to decrease in intensity, some of the other patterns increase their intensity in a manner that causes the troubled pattern to increase in intensity again

- **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 [16], 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 / credit assignment**. 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.

Next, for a variety of reasons outlined in [15] it becomes appealing to hypothesize that the network of patterns in an intelligent 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. Among many possible ways to hierarchically organize a set of patterns, the one used should be one that causes hierarchically nearby patterns to have many meaningful heterarchical connections; and of course, there should be a tendency to search for heterarchical connections among hierarchically nearby patterns.
- Self structure. A portion of the network of patterns forms into an approximate image of the overall network of patterns.

CogPrime has not been directly derived from these philosophical principles; rather, it has been created via beginning with a combination of human cognitive psychology and computer science algorithms and structures, and then shaping this combination so as to yield a system that appears likely to be conformant with these philosophical principles, as well as being computationally feasible on current hardware and containing cognitive structures and dynamics roughly homologous to the key human ones. The success of CogPrime as a design will depend largely on whether these high-level structures and dynamics can be made to emerge from the synergetic interaction of CogPrime 's representation and algorithms, when they are utilized to control an appropriate agent in an appropriate environment. The extended treatment of CogPrime given in [3], takes care to specifically elaborate how each of these abstract concepts arises concretely from CogPrime 's structures and algorithms.

IV. High-Level Architecture of CogPrime

Figure 1 depicts the high-level architecture of CogPrime. A key underlying principle is: **the use of multiple cognitive processes associated with multiple types of memory to enable an intelligent agent to execute the procedures that it believes have the best probability of working toward its goals in its current context**. In a robot preschool context, for example, the top-level goals would be everyday things such as pleasing the teacher, learning new information and skills, and protecting the robot's body.

It is interesting to compare these diagrams to the integrative human cognitive architecture diagram given in [17], which is intended to compactly overview the structure of human cognition as currently understood. The main difference is that the CogPrime diagrams commit to specific structures (e.g. knowledge representations) and processes, whereas the generic integrative architecture diagram refers merely to types of structures and processes. For instance, the integrative diagram refers generally to declarative knowledge and learning, whereas the CogPrime diagram refers to PLN, as a specific system for reasoning and learning about declarative knowledge. In [3] a table is provided articulating the key connections between the components of the CogPrime diagram and the well-known human cognitive structures/processes represented in integrative diagram, thus indicating the general cognitive functions instantiated by each of the CogPrime components.

V. Representation and Memory

A. Local and Global Knowledge Representation

One of the biggest decisions to make in designing an AGI system is how the system should represent knowledge. Naturally any advanced AGI system is going to synthesize a lot of its own knowledge representations for handling particular sorts of knowledge – but still, an AGI design typically makes *at least* some sort of commitment about the category of knowledge representation mechanisms toward which the AGI system will be biased.

OpenCog's knowledge representation mechanisms are all based fundamentally on *networks*. The view of mind as network is implicit in the patternist philosophy, because every pattern can be viewed as a pattern *in* something, or a pattern of arrangement *of* something – thus a pattern is always viewable as a relation between two or more things. A collection of patterns is thus a pattern-network. Knowledge of all kinds may be given network representations; and cognitive processes may be represented as networks also, for instance via representing them as programs, which may be represented as trees or graphs in various standard ways. The emergent patterns arising in an intelligence as it develops may be viewed as a pattern network in themselves; and the relations between an embodied mind and its physical and social environment may be viewed in terms of ecological and social networks.

The two major supercategories of knowledge representation systems are *local* (also called *explicit*) and *global* (also called *implicit*) systems, with a hybrid category we refer to as *glocal*

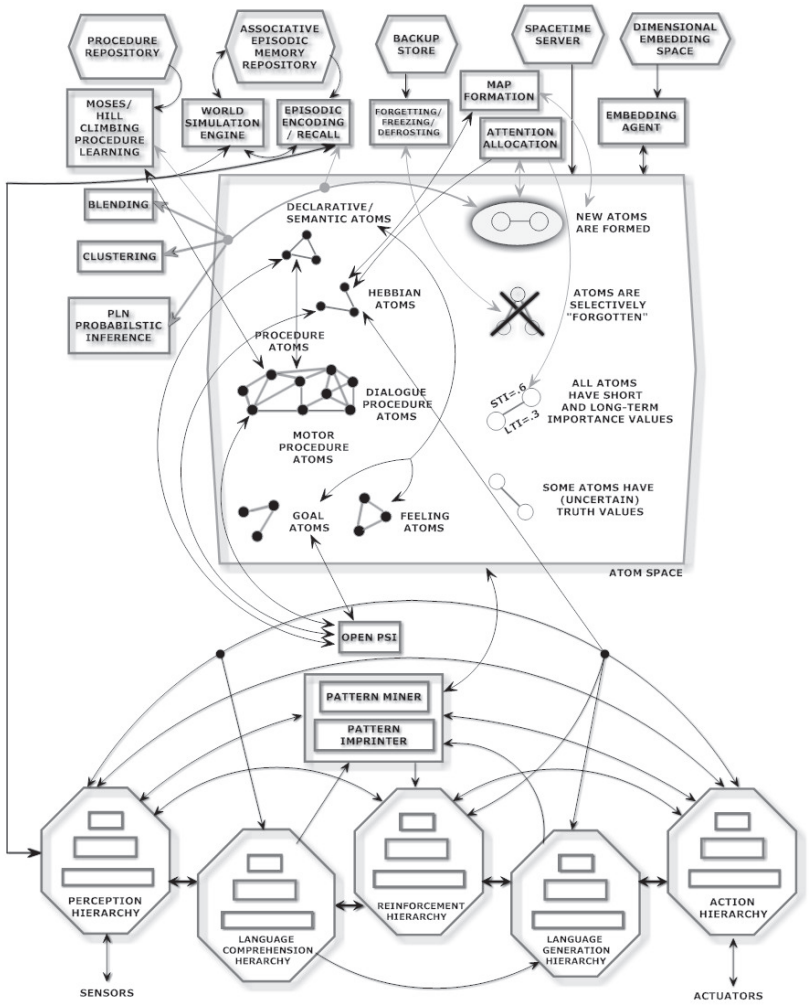

Fig. 1. **High-Level Architecture of CogPrime.** This is a conceptual depiction, not a detailed flowchart (which would be too complex for a single image).

that combines both of these. In a local system, each piece of knowledge is stored using a small percentage of cognitive system elements; in a global system, each piece of knowledge is stored using a particular pattern of arrangement, activation, etc. of a large percentage of cognitive system elements; in a glocal system, the two approaches are used together. All three of these knowledge representation types may be realized using networks. In CogPrime, all three are realized using the *same* (Atomspace) network.

B. Memory Types and Associated Cognitive Processes in CogPrime

CogPrime relies on multiple memory types and, as discussed above, is founded on the premise that the right course in architecting a pragmatic, roughly human-like AGI system is to handle different types of memory differently in terms of both structure and dynamics.

CogPrime's memory types are the declarative, procedural, sensory, and episodic memory types that are widely discussed in cognitive neuroscience [18], plus attentional memory for allocating system resources generically, and intentional memory for allocating system resources in a goal-directed way. Table I overviews these memory types, giving key references and indicating the corresponding cognitive processes, and also indicating which of the generic patternist cognitive dynamics each cognitive process corresponds to (pattern creation, association, etc.).

In terms of patternist cognitive theory, the multiple types of memory in CogPrime should be considered as specialized ways of storing particular types of pattern, optimized for spacetime efficiency. The cognitive processes associated with a certain type of memory deal with creating and recognizing patterns of the type for which the memory is specialized. While in principle all the different sorts of pattern could be handled in a unified memory and processing architecture, the sort of specialization used in CogPrime is necessary in order to achieve acceptable efficient general intelligence using currently available computational resources. And as we have argued in detail in [8], efficiency is not a side-issue but rather the essence of real-world AGI (since as Hutter has

shown, if one casts efficiency aside, arbitrary levels of general intelligence can be achieved via a trivially simple program).

The essence of the CogPrime design lies in the way the structures and processes associated with each type of memory are designed to work together in a closely coupled way, yielding cooperative intelligence going beyond what could be achieved by an architecture merely containing the same structures and processes in separate "black boxes."

The inter-cognitive-process interactions in OpenCog are designed so that conversion between different types of memory is possible, though sometimes computationally costly (e.g. an item of declarative knowledge may with some effort be interpreted procedurally or episodically, etc.); and so that when a learning process concerned centrally with one type of memory encounters a situation where it learns very slowly, it can often resolve the issue by converting some of the relevant knowledge into a different type of memory: i.e. **cognitive synergy**.

VI. Key Claims

Pulling together the conceptual and architectural points made above, we now summarize some of the key claims regarding human-level AGI that underly the CogPrime design. In essence this is a list of claims such that, if the reader accepts these claims, they should probably accept that the CogPrime approach to AGI is a viable one:

1) General intelligence (at the human level and ultimately beyond) can be achieved via creating a computational system that uses much of its resources seeking to achieve its goals, via using perception and memory to predict which actions will achieve its goals in the contexts in which it finds itself.
2) To achieve general intelligence in the context of human-intelligence-friendly environments and goals using feasible computational resources, it's important that an AGI system can handle different kinds of memory (declarative, procedural, episodic, sensory, intentional, attentional) in customized but interoperable ways.
3) Cognitive synergy: It's important that the cognitive processes associated with different kinds of memory can appeal to each other for assistance in overcoming bottlenecks in a manner that enables each cognitive process to act in a manner that is sensitive to the particularities of each others' internal representations, and that doesn't impose unreasonable delays on the overall cognitive dynamics.
4) As a general principle, neither purely localized nor purely global memory is sufficient for general intelligence under feasible computational resources; "glocal" memory will be required.
5) To achieve human-like general intelligence, it's important for an intelligent agent to have sensory data and motoric affordances that roughly emulate those available to humans. We don't know exactly how close this emulation needs to be, which means that our AGI systems and platforms need to support fairly flexible experimentation with virtual-world and/or robotic infrastructures.
6) To work toward adult human-level, roughly human-like general intelligence, one fairly easily comprehensible path is to use environments and goals reminiscent of human childhood, and seek to advance one's AGI system along a path roughly comparable to that followed by human children.
7) It is most effective to teach an AGI system aimed at roughly human-like general intelligence via a mix of spontaneous learning and explicit instruction, and to instruct it via a combination of imitation, reinforcement and correction, and a combination of linguistic and nonlinguistic instruction.
8) One effective approach to teaching an AGI system human language is to supply it with some in-built linguistic facility, in the form of rule-based and statistical-linguistics-based NLP systems, and then allow it to improve and revise this facility based on experience.
9) An AGI system with adequate mechanisms for handling the key types of knowledge mentioned (in item 2) above, and the capability to explicitly recognize large-scale patterns in itself, should, **upon sustained interaction with an appropriate environment in pursuit of appropriate goals**, emerge a variety of complex structures in its internal knowledge network, including interlocking hierarchical and heterarchical networks, and networks modeling itself and others.

Further, given the strengths and weaknesses of current and near-future digital computers,

1) a (loosely) neural-symbolic network is a good representation for directly storing many kinds of memory, and interfacing between those that it doesn't store directly;
2) Uncertain logic is a good way to handle declarative knowledge. To deal with the problems facing a human-level AGI, an uncertain logic must integrate imprecise probability and fuzziness with a broad scope of logical constructs. PLN is one good realization.
3) Programs are a good way to represent procedures (both cognitive and physical-action, but perhaps not including low-level motor-control procedures).
4) Evolutionary program learning is a good way to handle difficult program learning problems. Probabilistic learning on normalized programs is one effective approach to evolutionary program learning. MOSES is one good realization of this approach.
5) Activation spreading and Hebbian learning comprise a reasonable way to handle attentional knowledge (though other approaches, with greater overhead cost, may provide better accuracy and may be appropriate in some situations). Artificial economics (e.g. ECAN) is an effective approach to activation spreading and Hebbian learning in the context of neural-symbolic networks. A reasonable trade-off between comprehensiveness and efficiency is to focus on two kinds of attention: processor

Memory Type	Specific Cognitive Processes	General Cognitive Functions
Declarative	Probabilistic Logic Networks (PLN) [19]; conceptual blending [20]	pattern creation
Procedural	MOSES (a novel probabilistic evolutionary program learning algorithm) [9]	pattern creation
Episodic	internal simulation engine [10]	association, pattern creation
Attentional	Economic Attention Networks (ECAN) [21]	association, credit assignment
Intentional	probabilistic goal hierarchy refined by PLN and ECAN, structured according to the OpenPsi motivational framework (modeled on MicroPsi [22])	credit assignment, pattern creation
Sensory	In CogBot, this will be supplied by the DeSTIN component	association, attention allocation, pattern creation, credit assignment

TABLE I
MEMORY TYPES AND COGNITIVE PROCESSES IN COGPRIME. THE THIRD COLUMN INDICATES THE GENERAL COGNITIVE FUNCTION THAT EACH SPECIFIC COGNITIVE PROCESS CARRIES OUT, ACCORDING TO THE PATTERNIST THEORY OF COGNITION.

attention (represented in CogPrime by ShortTermImportance) and memory attention (represented in CogPrime by LongTermImportance).

6) Simulation is a good way to handle episodic knowledge (remembered and imagined). Running an internal world simulation engine is an effective way to handle simulation.
7) Hybridization of one's integrative neural-symbolic system with a spatiotemporally hierarchical deep learning system is an effective way to handle representation and learning of low-level sensorimotor knowledge. DeSTIN is one example of a deep learning system of this nature that can be effective in this context.
8) One effective way to handle goals is to represent them declaratively, and allocate attention among them economically. CogPrime 's PLN/ECAN based framework for handling intentional knowledge is one good realization.
9) It is important for an intelligent system to have some way of recognizing large-scale patterns in itself, and then embodying these patterns as new, localized knowledge items in its memory (a dynamic called the "cognitive equation" in [23]),. Given the use of a neural-symbolic network for knowledge representation, a graph-mining based "map formation" heuristic is one good way to do this.
10) Occam's Razor: Intelligence is closely tied to the creation of procedures that achieve goals in environments *in the simplest possible way*. Each of an AGI system's cognitive algorithms should embody a simplicity bias in some explicit or implicit form.
11) Once sufficiently advanced, an AGI system with a logic-based declarative knowledge approach and a program-learning-based procedural knowledge approach should be able to radically self-improve via a variety of methods.

Naturally, if some of these claims should prove false, then some portions of the CogPrime design may prove inadequate; but it may well be possibel to replace or repair them in a way consistent with the remainder of the architecture. The OpenCog framework is highly modular and supports experimentation with various approaches.

A. Conclusion

What we have sought to do in this brief review is:

- to roughly indicate a theoretical perspective on general intelligence, according to which the creation of a human-level AGI doesn't require anything *that* extraordinary, but "merely" an appropriate combination of closely interoperating algorithms operating on an appropriate multi-type memory system, utilized to enable a system in an appropriate body and environment to figure out how to achieve its given goals
- to roughly describe a software design (CogPrime) that, according to this somewhat mundane but theoretically quite well grounded vision of general intelligence, appears likely (according to a combination of rigorous and heuristic arguments) to be able to lead to human-level AGI using feasible computational resources
- to enumerate some of the key ideas and claims underlying this design

CogPrime is not an ad hoc software design, but a complex system created according to a network of principles regarding the overall systematic nature of mind, and the operation and interaction of the different parts of human-like minds. However, we wish to stress that *not all of the underlying arguments and ideas need to be 100% correct in order for the project to succeed*. We feel the quest to create AGI is a mix of theory, engineering, and scientific and unscientific experimentation. If the current CogPrime design turns out to have significant shortcomings, yet still brings us a significant percentage of the way toward human-level AGI, the results obtained along the path will very likely give us clues about

how to tweak the design to more effectively get the rest of the way there. And the OpenCog platform is extremely flexible and extensible, rather than being tied to the particular details of the CogPrime design. While we do have faith that the CogPrime design as described here has human-level AGI potential, we are also pleased to have a development strategy and implementation platform that will allow us to modify and improve the design in accordance with whatever suggestions are made by our ongoing experimentation.

Waxing historical for a moment, we note that many great achievements in history have seemed more magical before their first achievement than afterwards. Powered flight and spaceflight are the most obvious examples, but there are many others such as mobile telephony, prosthetic limbs, electronically deliverable books, robotic factory workers, and so on. We now even have wireless transmission of power (one can recharge cellphones via wifi), though not yet as ambitiously as Tesla envisioned. We very strongly suspect that human-level AGI is in the same category as these various examples: an exciting and amazing achievement, which however is achievable via systematic and careful application of fairly mundane principles. We believe computationally feasible human-level intelligence is both *complicated* (involving many interoperating parts, each sophisticated in their own right) and *complex* (in the sense of involving many emergent dynamics and structures whose details are not easily predictable based on the parts of the system) ... but that neither the complication nor the complexity is an obstacle to engineering human-level AGI.

In our view, what is needed to create human-level AGI is not a new scientific breakthrough, nor a miracle, but "merely" a sustained effort over a number of years by a moderate-sized team of appropriately-trained professionals, completing the implementation of an adequate design – such as the one described in [3] and roughly sketched here – and then parenting and educating the resulting implemented system.

REFERENCES

[1] B. Goertzel and C. Pennachin, *Artificial General Intelligence*. Springer, 2005.

[2] B. Goertzel, H. Pinto, C. Pennachin, and I. F. Goertzel, "Using dependency parsing and probabilistic inference to extract relationships between genes, proteins and malignancies implicit among multiple biomedical research abstracts," in *Proc. of Bio-NLP 2006*, 2006.

[3] B. Goertzel, C. Pennachin, and N. Geisweiller, *Building Better Minds: Engineering Beneficial General Intelligence*. In preparation, 2013.

[4] B. Goertzel, R. Lian, H. de Garis, S. Chen, and I. Arel, "World survey of artificial brains, part ii: Biologically inspired cognitive architectures," *Neurocomputing*, Apr. 2010.

[5] H. De Garis, B. Goertzel, R. Lian, H. de Garis, and S. Chen, "World survey of artificial brains, part i: Brain simulations," *Neurocomputing*, 2010.

[6] R. Kurzweil, *The Singularity is Near*, 2006.

[7] S. Legg and M. Hutter, "A collection of definitions of intelligence." IOS, 2007.

[8] B. Goertzel, "Toward a formal definition of real-world general intelligence," 2010.

[9] M. Looks, *Competent Program Evolution*. PhD Thesis, Computer Science Department, Washington University, 2006.

[10] B. Goertzel and C. P. Et Al, "An integrative methodology for teaching embodied non-linguistic agents, applied to virtual animals in second life," in *Proc.of the First Conf. on AGI*. IOS Press, 2008.

[11] B. Goertzel, J. Pitt, Z. Cai, J. Wigmore, D. Huang, N. Geisweiller, R. Lian, and G. Yu, "Integrative general intelligence for controlling game ai in a minecraft-like environment," in *Proc. of BICA 2011*, 2011.

[12] B. Goertzel and H. de Garis, "Xia-man: An extensible, integrative architecture for intelligent humanoid robotics," pp. 86–90, 2008.

[13] B. Goertzel, "Perception processing for general intelligence, part ii: Bridging the symbolic/subsymbolic gap." (in preparation).

[14] I. Arel, D. Rose, and R. Coop, "Destin: A scalable deep learning architecture with application to high-dimensional robust pattern recognition." *Proc. AAAI Workshop on Biologically Inspired Cognitive Architectures*, 2009.

[15] B. Goertzel, *The Hidden Pattern*. Brown Walker, 2006.

[16] C. Peirce, *Collected papers: Volume V. Pragmatism and pragmaticism*. Harvard University Press. Cambridge MA., 1934.

[17] B. Goertzel, M. Ikle', and J. Wigmore, "The architecture of human-like general intelligence," in *Foundations of Artificial General Intelligence*, 2012.

[18] E. Tulving and R. Craik, *The Oxford Handbook of Memory*. Oxford U. Press, 2005.

[19] B. Goertzel, I. G. M. Ikl, and A. Heljakka, *Probabilistic Logic Networks*. Springer, 2008.

[20] G. Fauconnier and M. Turner, *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic, 2002.

[21] B. Goertzel, J. Pitt, M. Ikle, C. Pennachin, and R. Liu, "Glocal memory: a design principle for artificial brains and minds," *Neurocomputing*, Apr. 2010.

[22] J. Bach, *Principles of Synthetic Intelligence*. Oxford University Press, 2009.

[23] B. Goertzel, *Chaotic Logic*. Plenum, 1994.