International Journal of Machine Consciousness
Vol. 3, No. 1 (2011) 19–53
© World Scientific Publishing Company
DOI: 10.1142/S1793843011000601

HYPERSET MODELS OF SELF, WILL AND REFLECTIVE CONSCIOUSNESS

BEN GOERTZEL

Novamente LLC 1405 Bernerd Place,
Rockville MD 20851, USA
ben@goertzel.org

A novel theory of reflective consciousness, will and self is presented, based on modeling each of these entities using self-referential mathematical structures called hypersets. Pattern theory is used to argue that these exotic mathematical structures may meaningfully be considered as parts of the minds of physical systems, even finite computational systems. The hyperset models presented are hypothesized to occur as patterns within the "moving bubble of attention" of the human brain and any roughly human-mind-like AI system. These ideas appear to be compatible with both panpsychist and materialist views of consciousness, and probably other views as well. Their relationship with the CogPrime AI design and its implementation in the OpenCog software framework is elucidated in detail.

Keywords: Consciousness; conscious reflection; self; will; hypersets.

1. Introduction

What is consciousness; what is conscious reflection? What is the conscious will? What is the self; what is self-consciousness?

David Chalmers [1997] has famously declared that the "hard problem" of consciousness is understanding the fundamental nature of the connection between subjective experiences and the physical structures and dynamics associated with these. We do not deal with the "hard problem" here, but rather address the "easier" question: If one does assume the existence of correlations between experiences and physical structures and dynamics, then *which sorts of physical structures and dynamics correspond with which sorts of experiences*?

Pointing to specific regions or dynamic phenomena in the brain and associating them with aspects of human experience is interesting but does not answer the question that concerns us. What we are interested in here are the *abstract structures* occurring in the physical world, corresponding with particular types of subjective experience. Specifically, we want to know which abstract structures correspond to the experiences of "free will", reflective consciousness, and the phenomenal self [Metzinger, 2004]. We will propose some novel answers to these questions, using some mathematics not usually discussed in this context (hypersets). In spite of the use of advanced mathematics the overall treatment will be relatively informal: the goal here is to put forth a set of new ideas, which may then be dissected, explored and applied in much more detail in later papers.

The main ideas presented here make sense under various different philosophies of consciousness. However, for sake of simplicity and concreteness, we will discuss them here in the context of only two such philosophies: panpsychism and materialism, considered roughly as follows:

- The reader may see *The Hidden Pattern* [Goertzel, 2006a] for details on our own particular flavor of panpsychism; but in brief, we view a certain amount of consciousness as inherent in everything, and then understand different entities as manifesting universal consciousness in different sorts of ways. In this view, free will, reflective consciousness and phenomenal self correspond to different manifestations of universal consciousness.
- On the other hand, by materialism we mean the simple hypothesis that experiences *are* the physical structures and dynamics that correspond to them — i.e., that there is the physical world and nothing else. Dennett's perspective in Dennett [1993] is a paradigm case of this view.

We discuss "subjective experiences" at several points in the following. The panpsychist and the materialist may interpret this phrase differently. The panpsychist will interpret these references as indicating actual subjective experiences. On the other hand, the materialist reader may interpret all of our references to "subjective experiences" as meaning "situations corresponding to reported subjective experience". In the latter view, our investigation is interpreted as a study of which abstract structures correspond to states of mind where intelligences *report* experiences of free will, reflective consciousness and the phenomenal self.

Our core hypothesis here is that the abstract structures corresponding to free will, reflective consciousness and phenomenal self are effectively modeled using the mathematics of *hypersets* — where "hyperset" is an informal term used to refer to a mathematical set defined under a set of axioms that allows circular membership structures.

While the specific ideas presented here are novel — and in fact I have not found any prior reference to hypersets as a mode of consciousness at all — the idea of analyzing consciousness and related structures in terms of infinite recursions and non-foundational structures has occurred before, for instance in the works of Douglas Hofstadter [1979], G. Spencer-Brown [1967], Louis Kauffmann [n.d.] and Francisco Varela [1979]. None of these works uses hypersets in particular; but a more important difference is that none of them attempts to deal with particular psychological phenomena in terms of correlation, causation, pattern theory or similar concepts; they essentially stop at the point of noting the presence of a formalizable pattern of infinite recursion in reflective consciousness. Varela [1979] does venture into practical psychology via porting some of R. D. Laing's psychosocial "knots" [1972] into a formal non-foundational language; but this is a very specialized exercise that does not involve modeling general psychological structures or processes. Situation semantics [Barwise, 1989] does analyze various commonsense concepts and relationships using hypersets; however, it does not address issues of subjective experience explicitly, and does not present formal treatments of the phenomena considered here.

As yet we have not validated the models suggested here in any formal way, so they are presented only as interesting and intuitively appealing hypotheses. At the end of the paper, we will briefly outline ways in which they could be tested in future via analysis of neuroimaging data and execution traces of AI systems. Due to the potential for future empirical validation, the ideas presented here may be considered to lie on the borderline between philosophy and science.

1.1. *What are hypersets?*

What are these things called hypersets, which we posit as models of consciousness and related phenomena?

In the standard axiomatizations of set theory, such as Zermelo−Frankel set theory Devlin [1984], there is an axiom called the Axiom of Foundation, which implies that no set can contain itself as a member. That is, it implies that all sets are "well founded" — they are built up from other sets, which in turn are built up from other sets, etc., ultimately being built up from the empty set or from atomic elements. The hierarchy via which sets are built from other sets may be infinite (according to the usual Axiom of Infinity), but it goes in only one direction — if set A is built from set B (or from some other set built from set B), then set B cannot be built from set A (or from any other set built from set A).

However, since very shortly after the Axiom of Foundation was formulated, there have been various alternative axiomatizations which allow "non-well-founded" sets (aka hypersets), i.e., sets that *can* contain themselves as members, or have more complex circular membership structures. Hyperset theory is generally formulated as an extension of classical set theory rather than a replacement — i.e., the well-founded sets within a hyperset domain conform to classical set theory. In recent decades the theory of non-well-founded sets has been applied in computer science (e.g., process algebra [Aczel, 1984]), linguistics and natural language semantics (situation theory [Barwise, 1989]), philosophy (work on the Liar Paradox [Barwise and Etchemendy, 1989]), and other areas.

For instance, in hyperset theory you can have

$$A = \{A\}$$

$$A = \{B, \{A\}\}$$

and so forth. Using hypersets you can have functions that take themselves as arguments, and many other interesting phenomena that are not permitted by the standard axioms of set theory. The main work of this paper is to suggest specific models of free will, reflective consciousness and phenomenal self in terms of hyperset mathematics.

The reason the Axiom of Foundation was originally introduced was to avoid paradoxes like the Russell Set (the set of all sets that contain themselves). None of these variant set theories allow all possible circular membership structures; but they allow restricted sets of such, sculpted to avoid problems like the Russell Paradox.

One currently popular form of hyperset theory is obtained by replacing the Axiom of Foundation with the Anti-Foundation Axiom (AFA) which, roughly speaking, permits circular membership structures that map onto graphs in a certain way. All the hypersets discussed here are easily observed to be allowable under the AFA (according to the Solution Lemma stated in Aczel [1988]).

Specifically, the AFA uses the notion of an *accessible pointed graph* — a directed graph with a distinguished element (the "root") such that for any node in the graph there is at least one path in the directed graph from the root to that node. The AFA states that every accessible pointed graph corresponds to a unique set. For example, the graph consisting of a single vertex with a loop corresponds to a set which contains only itself as element.

1.2. *The panpsychist perspective*

The hyperset models of consciousness and related phenomena presented here are not intrinsically tied to any specific philosophy of consciousness. For example,

- If one adopts a materialist perspective on consciousness, then it will one day be possible to test the present ideas, by asking whether the posited hyperset structures really are detectable in those intelligent systems that self-report the experiences posited to correspond with them.
- If one adopts a panpsychist perspective, then the correlation between the posited structures and the posited subjective experiences becomes something to be validated via a combination of scientific analysis and personal introspection.

A related observation is that the treatment given here mixes up empirical and introspective matters in a fairly free and easy way. This is not done without premeditation, and merits brief discussion:

- From a materialist point of view, this mixture is not really problematic, since introspections may be interpreted as "reported introspections".
- From a panpsychist point of view, the matter is subtler. We suspect that various issues related to consciousness may be more tractable within a future discipline, yet to be fleshed out, that combines aspects of contemporary science with introspective aspects. Francisco Varela was pushing toward such a discipline in Shear and Varela [2001] and Thompson [2001]. While the dimensions of this hypothesized future discipline are not yet clear, we suspect that it will allow intermixture of empirical and experiential aspects in the manner pursued here.

But even though the ideas presented here are not logically tied to the panpsychist perspective, they did emerge initially from this perspective, and so in the following subsection, before starting the main line of argument of the paper, we will briefly review panpsychism as we understand it.

1.3. *Panpsychism*

Panpsychism occurs in various forms, but in the broad sense it refers simply to the idea that mind is a fundamental feature of the universe and each of its parts, rather than something that is the exclusive property of specific kinds of systems like humans, other higher animals, intelligent computer programs, etc. [Seager and Allen-Hermanson, 2010].

Though not a common view in contemporary Western society, philosophy or science, panpsychism does have a long history in historical Western philosophy, encompassing thinkers like Leibniz, James, Whitehead, Russell, Fechner and Spinoza. A host of recent books treat the topic, including Skrbina's *Mind that Abides*: *Panpsychism in the New Millienium* [2009] and Strawson *et al.*'s *Consciousness and its Place in Nature* [2006].

Panpsychism also has a long and rich history in Eastern philosophy, e.g., the modern Vedantic thinker Swami Krishnananda [2010] observes:

> *The Vedanta philosophy concludes that matter also is a phase of consciousness and objects of knowledge embody in themselves a hidden potential of consciousness which is also the Self of the perceiving subject*, *enabling experience in the subject. The subject-consciousness* (*Vishayi-chaitanya*) *is in a larger dimension of its own being as universality and all-pervadingness beholds itself in the object-consciousness* (*Vishaya-chaitanya*), *thereby reducing all possible experience to a degree of universal consciousness. Experience is neither purely subjective nor entirely objective*; *experience is caused by the universal element inherent in both the subject and the object*, *linking the two terms of the relation together and yet transcending both the subject and the object because of its universality.*

Advocates of panpsychism point out that alternative theories of mind and consciousness are riddled with problems and inconsistencies, whereas panpsychism is simple and coherent, its only "problem" being that it disagrees with the intuition of many modern Western folk. Most current theories of consciousness involve mind and awareness somehow emerging out of non-sentient matter, which is conceptually problematic. Philosopher Galen Strawson has recently lamented the basic senselessness of the notion that mental experience can emerge from a wholly non-mental, non-experiential substrate: "I think it is very, very hard to understand what it is supposed to involve. I think that it is incoherent, in fact …" [Strawson, 2006].

Dualist theories in which the mind-realm and the matter-realm are separate but communicating also run into difficulties, e.g., the problem that (put crudely) the mind-realm must be utterly undetectable via science or else in effect it becomes part of the matter-realm. Panpsychism holds that everything in the world has mental extent, similar to how it has spatial and temporal extent, which is a simple proposal that does not give rise to any conceptual contradictions.

Some have objected to panpsychism due to the apparent lack of evidence that the fundamental entities of the physical world possess any mentalistic properties. However, this lack of evidence may easily be attributed to our poor observational skills. By analogy, humans cannot directly detect the gravitational properties of small objects, but this does not render such properties nonexistent. And in appropriate states of consciousness, humans *can* directly apprehend the consciousness of objects like rocks, chairs or particles, a fact driven home forcefully by Aldous Huxley in *The Perennial Philosophy* [1990].

Panpsychism is not without its difficulties, e.g., the combination problem, first raised by William James — which in essence wonders: if everything is conscious, how does the consciousness of a whole relate to the consciousnesses of its parts? How does the brain's consciousness come out of the consciousnesses of its component neurons, for example? [James, 1950].

But this does not seem a problem on the order of "how does consciousness emerge from non-conscious matter", it seems more a technical issue. A large variety of qualitatively different part-whole relationships may exist, as physicists have noted in the last century. Quantum mechanics has made clear that systems are not simply the sum of their parts but can sometimes exhibit properties that go beyond those of the parts and which cannot be detected by examining the parts in isolation. And black hole physics has shown us the possibility of wholes (black holes) that totally lose most of the properties possessed by their parts and render the parts inaccessible (a black hole has only the properties of mass, charge and spin, regardless of the other properties possessed by the objects that combined to form the black hole). The nature of part-whole relationships in panpsychism certainly bears further study, but merely appears subtle, not incoherent. And the emergent, holistic aspect of consciousness is well known in Eastern thought, e.g., Swami Krishnananda says that

> *The three states of waking, dream and sleep, through which we pass in our daily experience, differ from one another, and yet a single consciousness connects them, enabling the individual to experience an identity even in the otherwise differentiatedness of these states. Since consciousness links the three states into a singleness of experience, it is immanent in them and yet transcends them, not capable of identity with any of them.*

In short, the panpsychist view of consciousness has a long history in both Eastern and Western philosophy, and has no glaring conceptual problems associated with it, the only real "issue" with it being that most people in contemporary Western cultures find it counterintuitive. I have found it a useful guide for thinking about the mind, perhaps largely because it does not contain any confusing inconsistencies or incoherences that "get in the way" of analyzing other issues such as the ones considered in the rest of this paper, reflective consciousness, self and will.

2. Patterns, Correlations and Experience

One of the foundations of the ideas presented here is the hypothesis, made in *The Hidden Pattern*, that the subjective experience of being conscious of some entity X, is correlated with the presence of a very intense pattern in one's overall mind-state, corresponding to X. This simple idea is also the essence of neuroscientist Susan Greenfield's theory of consciousness [2001] (but in her theory, "overall mind-state" is replaced with "brain-state"), and has much deeper historical roots in philosophy of mind which we shall not venture to unravel here.

This observation relates to the idea of "moving bubbles of awareness" in intelligent systems. If an intelligent system consists of multiple processing or data elements, and during each (sufficiently long) interval of time some of these elements get much more attention than others, then one may view the system as having a certain "attentional focus" during each interval. The attentional focus is itself a significant pattern in the system (the pattern being "these elements habitually get more processor and memory", roughly speaking). As the attentional focus shifts over time one has a "moving bubble of pattern" which then corresponds experientially to a "moving bubble of awareness".

This notion of a "moving bubble of awareness" ties in very closely to global workspace theory [Baars and Franklin, 2009], a cognitive theory that has broad support from neuroscience and cognitive science and has also served as the motivation for Stan Franklin's LIDA AI system [Baars and Franklin, 2009]. The global workspace theory views the mind as consisting of a large population of small, specialized processes, a society of agents. These agents organize themselves into coalitions, and coalitions that are relevant to contextually novel phenomena, or contextually important goals, are pulled into the global workspace (which is identified with consciousness). This workspace broadcasts the message of the coalition to all the unconscious agents, and recruits other agents into consciousness. Various sorts of contexts — e.g., goal contexts, perceptual contexts, conceptual contexts and cultural contexts — play a role in determining which coalitions are relevant, and form the unconscious background of the conscious global workspace. New perceptions are often, but not necessarily, pushed into the workspace. Some of the agents in the global workspace are concerned with action selection, i.e., with controlling and passing parameters to a population of possible actions. The contents of the workspace at any given time have a certain cohesiveness and interdependency, the so-called unity of consciousness. In essence the contents of the global workspace form a moving bubble of attention or awareness.

In the OpenCog AI system [Goertzel, 2009a] that my colleagues and I have developed (and that will be discussed in more depth below), this moving bubble is achieved via economic attention network (ECAN) equations [Goertzel *et al.*, 2010a] that propagate virtual currency between nodes and links representing elements of memories, so that the attentional focus consists of the wealthiest nodes and links. Figures 1 and 2 illustrate the existence and flow of attentional focus in OpenCog.

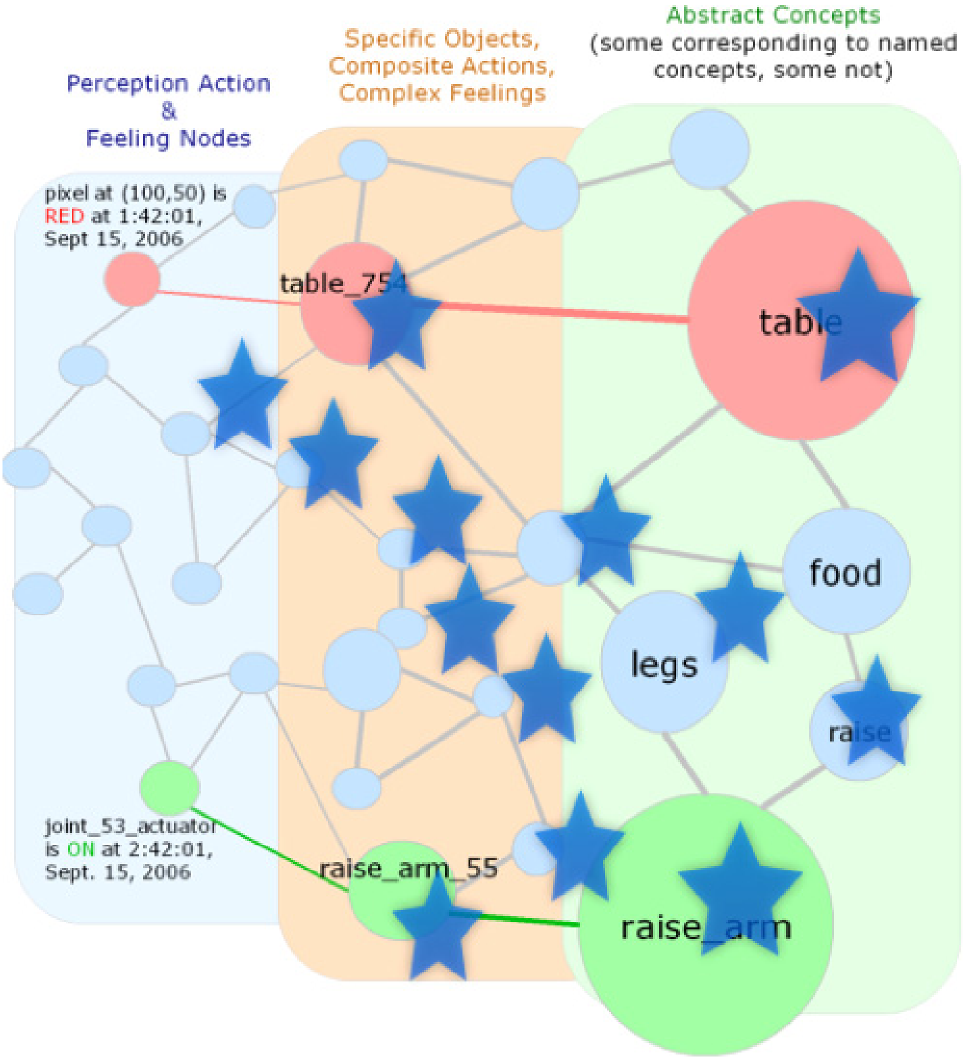

Fig. 1. Graphical depiction of the momentary bubble of attention in the memory of an OpenCog AI system. Circles and lines represent nodes and links in OpenCog's memory, and stars denote those nodes with a high level of attention (represented in OpenCog by the ShortTermImportance node variable) at the particular point in time.

On the other hand, in Hameroff's [2010] recent model of the brain, the brain's moving bubble of attention is achieved through dendro-dendritic connections and the emergent dendritic web.

In this broad perspective, self, free will and reflective consciousness are specific phenomena occurring *within* the moving bubble of awareness. They are specific ways of experiencing awareness, corresponding to certain abstract types of physical structures and dynamics, which we shall endeavor to identify here.

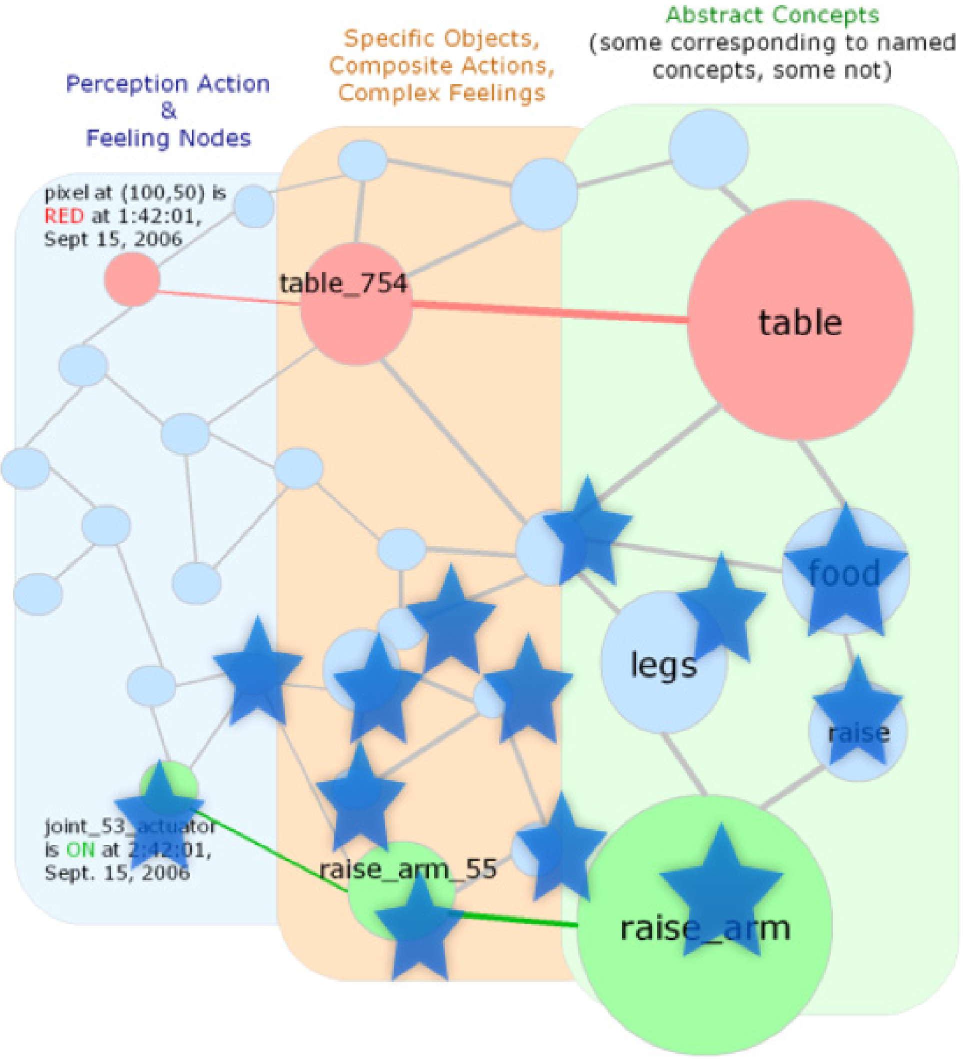
Perception Action & Feeling Nodes
Specific Objects, Composite Actions, Complex Feelings
Abstract Concepts (some corresponding to named concepts, some not)
pixel at (100,50) is RED at 1:42:01, Sept 15, 2006
table_754
table
food
legs
raise
joint_53_actuator is ON at 2:42:01, Sept. 15, 2006
raise_arm_55
raise_arm

Fig. 2. Graphical depiction of the momentary bubble of attention in the memory of an OpenCog AI system, a few moments after the bubble shown in Fig. 1, indicating the moving of the bubble of attention. Depictive conventions are the same as in Fig. 1. This shows an idealized situation where the declarative knowledge remains invariant from one moment to the next but only the focus of attention shifts. In reality both will evolve together.

Patternism and panpsychism fit very neatly together — according to this combination, one posits simply that **every pattern is conscious, but different sorts of patterns have different flavors of consciousness; and some patterns are more intensely conscious than others**. The appearance of reflective consciousness in humans and parrots but not roaches or rocks is not paradoxical or shocking — it means that the qualia associated with roaches and rocks do not have the particular property of deliberative reflection. And the patterns in a person's "unconscious" mind are understood to have their own flavor of consciousness — just as do the patterns in a bug — but not reflective, deliberative or verbal consciousness.

3. A Patternist Perspective on Mind and Consciousness

To more fully understand the view of consciousness hinted above, it is helpful to have a little background in the "patternist philosophy of mind" overall. Patternism a general approach to thinking about intelligent systems, 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.

Pattern as the basis of mind is not in itself a very novel idea; this concept is present, for instance, in the 19th-century philosophy of Charles Peirce [1935], in the writings of contemporary philosophers Daniel Dennett [1993] and Douglas Hofstadter [1979], in Benjamin Whorf's [1964] linguistic philosophy and Gregory Bateson's [1979] systems theory of mind and nature. Bateson spoke of the Metapattern: "that it is pattern which connects". In my prior writings on philosophy of mind, I have sought to pursue this theme more thoroughly than has been done before, and to articulate in detail how various aspects of human mind and mind in general can be well-understood by explicitly adopting a patternist perspective.[1]

In the patternist perspective, "pattern" is generally defined as "representation as something simpler". Thus, for example, if one measures simplicity in terms of bit-count, then a program compressing an image would be a pattern in that image. But if one uses a simplicity measure incorporating run-time as well as bit-count, then the compressed version may or may not be a pattern in the image, depending on how one's simplicity measure weights the two factors. This definition encompasses simple repeated patterns, but also much more complex ones. While pattern theory has typically been elaborated in the context of computational theory, it is not intrinsically tied to computation; rather, it can be developed in any context where there is a notion of "representation" or "production" and a way of measuring simplicity. One just needs to be able to assess the extent to which f represents or produces X, and then to compare the simplicity of f and X; and then one can assess whether f is a pattern in X. A formalization of this notion of pattern is given in [Goertzel, 2006a] and briefly summarized at the end of this section.

Next, in patternism the mind of an intelligent system is conceived as the (fuzzy) set of patterns in that system, and the set of patterns emergent between that system and other systems with which it interacts. The latter clause means that the patternist perspective is inclusive of notions of distributed intelligence [Hutchins, 1996]. Basically, the mind of a system is the fuzzy set of different simplifying representations of that system that may be adopted.

[1] In some prior writings the term "psynet model of mind" has been used to refer to the application of patternist philosophy to cognitive theory, but this term has been avoided in recent publications as it seemed to introduce more confusion than clarification.

Intelligence is conceived, similarly to in Marcus Hutter's [2005] recent work, as the ability to achieve complex goals in complex environments; where complexity itself may be defined as the possession of a rich variety of patterns. A mind is thus a collection of patterns that is associated with a persistent dynamical process that achieves highly-patterned goals in highly-patterned environments.

An additional hypothesis made within the patternist philosophy of mind is that reflection is critical to intelligence. This lets us conceive an intelligent system as a dynamical system that recognizes patterns in its environment and itself, as part of its quest to achieve complex goals.

While this approach is quite general, it is not vacuous; it gives a particular structure to the tasks of analyzing and synthesizing intelligent systems. About any would-be intelligent system, we are led to ask questions such as:

- How are patterns represented in the system? That is, how does the underlying infrastructure of the system give rise to the displaying of a particular pattern in the system's behavior?
- What kinds of patterns are most compactly represented within the system?
- What kinds of patterns are most simply learned?
- What learning processes are utilized for recognizing patterns?
- What mechanisms are used to give the system the ability to introspect (so that it can recognize patterns in itself)?

Of course, these same sorts of questions could be asked if one substituted the word "pattern" with other words like "knowledge" or "information". However, we have found that asking these questions in the context of pattern leads to more productive answers, avoiding unproductive byways and also tying in very nicely with the details of various existing formalisms and algorithms for knowledge representation and learning.

Among the many kinds of patterns in intelligent systems, *semiotic* patterns are particularly interesting ones. Peirce decomposed these into three categories:

- **iconic** patterns, which are patterns of contextually important internal similarity between two entities (e.g., an iconic pattern binds a picture of a person to that person)
- **indexical** patterns, which are patterns of spatiotemporal co-occurrence (e.g., an indexical pattern binds a wedding dress and a wedding)
- **symbolic** patterns, which are patterns indicating that two entities are often involved in the same relationships (e.g., a symbolic pattern between the number "5" (the symbol) and various sets of five objects (the entities that the symbol is taken to represent)).

Beyond semiotics, 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 is defined rigorously, but basically amounts to an intelligent agent 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 example, if an intelligent 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 alot 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 Goertzel [2006a], several primary dynamical principles may be posited, including:

- **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 inter-related 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 [1935] law of mind, 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.

And, for a variety of reasons outlined in Goertzel [2006a] 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.

3.1. *Appendix: Quantifying pattern*

In this subsection, we follow up the above informal overview of patternism with a brief review of the formalization of the notion of "pattern" given in Appendix 1 of [Goertzel, 2006a], with some minor additions. This formalization is what allows us to articulate the sense in which a hyperset can be considered a pattern in a physical system, even a finite system.

Definition 1. Given a metric space (M, d), and two functions $c : M \to [0, \infty]$ (the "simplicity measure") and $F : M \to M$ (the "production relationship"), we say that $\mathcal{P} \in M$ is a **pattern** in $X \in M$ to the degree

$$\iota_X^{\mathcal{P}} = \left((1 - d(F(\mathcal{P}), X)) \frac{c(X) - c(\mathcal{P})}{c(X)} \right)^+$$

This degree is called the **pattern intensity** of $\mathcal{P}$ in X.

For instance, if one wishes one may take c to denote algorithmic information measured on some reference Turing machine, and $F(X)$ to denote what appears on the second tape of a two-tape Turing machine t time-steps after placing X on its first tape. Other more naturalistic computational models are also possible here and are discussed extensively in Appendix 1 of [Goertzel, 2006a].

Definition 2. The **structure** of $X \in M$ is the fuzzy set St_X defined via the membership function

$$\chi_{St_X}(\mathcal{P}) = \iota_X^{\mathcal{P}}$$

This leads up to the formal definition of "mind" given in [Goertzel, 2006a]: the mind of X is the set of patterns associated with X. We can formalize this, for instance, by considering $\mathcal{P}$ to belong to the mind of X if it is a pattern in some Y that includes X. There are then two numbers to look at: $\iota_X^{\mathcal{P}}$ and $P(Y|X)$ (the percentage of Y that is also contained in X). To define the degree to which $\mathcal{P}$ belongs to the mind of X we can then combine these two numbers using some function f that is monotone increasing in both arguments. This highlights the somewhat arbitrary semantics of "of" in the phrase "the mind of X". Which of the patterns binding X to its environment are part of X's mind, and which are part of the world? This is not necessarily a good question, and the answer seems to depend on what perspective you choose, represented formally in the present framework by what combination function f you choose (for instance if $f(a,b) = a^r b^{2-r}$ then it depends on the choice of $0 < r < 1$).

Next, consider the case where the metric space M has a partial ordering $<$ on it; we may then define

Definition 3. $\mathcal{R} \in M$ is a **subpattern** in $X \in M$ to the degree

$$\kappa_X^{\mathcal{R}} = \frac{\int_{\mathcal{P}\in M} \text{true}(R < P)\, d\iota_X^{\mathcal{P}}}{\int_{\mathcal{P}\in M} d\iota_X^{\mathcal{P}}}$$

This degree is called the **subpattern intensity** of $\mathcal{P}$ in X.

Roughly speaking, the subpattern intensity measures the percentage of patterns in X that contain R (where "containment" is judged by the partial ordering $<$). But the percentage is measured using a weighted average, where each pattern is weighted by its intensity as a pattern in X. A subpattern may or may not be a pattern on its own. A nonpattern that happens to occur within many patterns may be an intense subpattern.

Whether the subpatterns in X are to be considered part of the "mind" of X is a somewhat superfluous question of semantics. Here we will extend the definition of mind given in [Goertzel, 2006a] to include subpatterns as well as patterns, because this makes it simpler to describe the relationship between hypersets and minds.

4. Hypersets as Patterns in Physical or Computational Systems

Hypersets are large infinite sets — they are certainly not computable — and so one might wonder if a hyperset model of consciousness supports Penrose [1996] and Hameroff's [1987] notion of consciousness as involving as-yet unknown physical dynamics involving uncomputable mathematics. However, this is not our perspective.

In the following we will present a number of particular hypersets and discuss their presence as patterns in intelligent systems. But this does not imply that we are positing intelligent systems to fundamentally *be* hypersets, in the sense that classical physics posits intelligent systems to be matter in $3 + 1$-dimensional space. Rather, we are positing that it is possible for hypersets to serve as *patterns* in physical systems, where the latter may be described in terms of classical or modern physics, or in terms of computation.

How is this possible? If a hyperset can *produce* a somewhat accurate model of a physical system, and is judged *simpler* than a detailed description of the physical system, then it may be a pattern in that system according to the definition of pattern given above.

Referring back to the above definition, define the metric space M to contain both hypersets and computer programs, and also tuples whose elements may be freely drawn from either of these classes. Define the partial order $<$ so that if X is an entry in a tuple T, then $X < T$.

Distance between two programs may be defined using the algorithmic information metric

$$d_I(A, B) = I(A|B) + I(B|A)$$

where $I(A|B)$ is the length of the shortest self-delimiting program for computing A given B [Chaitin, 2008]. Distance between two hypersets X and Y may be defined as

$$d_H(X, Y) = d_I(g(A), g(B))$$

where $g(A)$ is the graph (A's apg, in AFA lingo) picturing A's membership relationship. If A is a program and X is a hyperset, we may set $d(A, X) = \infty$.

Next, the production relation F may be defined to act on a (hyperset, program) pair $P = (X, A)$ via feeding the graph representing X (in some standard encoding) to A as an input. According to this production relation, P may be a pattern in the bit string $B = A(g(X))$; and since $X < P$, the hyperset X may be a subpattern in the bit string B.

It follows from the above that a hyperset can be part of the mind of a finite system described by a bit string, a computer program, or some other finite representation. But what sense does this make conceptually? Suppose that a finite system S contains entities of the form

$$\begin{aligned} C \\ G(C) \\ G(G(C)) \\ G(G(G(C))) \\ \ldots \end{aligned}$$

Then it may be effective to compute S using a (hyperset, program) pair containing the hyperset

$$X = G(X)$$

and a program that calculates the first k iterates of the hyperset. If so, then the hyperset $\{X = G(X)\}$ may be a subpattern in S. We will see some concrete examples of this in the following.

Whether one thing is a pattern in another depends not only on production but also on relative simplicity. So, if a system is studied by an observer who is able to judge some hypersets as simpler than some computational entities, then there is the possibility for hypersets to be subpatterns in computational entities, according to that observer. For such an observer, there is the possibility to model mental phenomena like will, self and reflective consciousness as hypersets, consistently with the conceptualization of mind as pattern.

5. A Hyperset Model of Reflective Consciousness

Whatever your view of the ultimate nature of consciousness, you probably agree that different entities in the universe manifest different *kinds* of consciousness or "awareness". Worms are aware in a different way than rocks; and dogs, pigs, pigeons and people are aware in a different way from worms. In [Goertzel, 1994] it is argued that hypersets can be used to model the sense in which the latter beasts are conscious whereas worms are not, i.e., what might be called "reflective consciousness".

We shall begin with the old cliché that

$$\text{Consciousness is consciousness of consciousness}$$

Note that this is nicely approximated by the series

$$\begin{array}{r} A \\ \text{Consciousness of } A \\ \text{Consciousness of consciousness of } A \\ \cdots \end{array}$$

This is quite conceptually nice, but does not really serve as a definition or precise characterization of consciousness. Even if one replaces it with

$$\text{Reflective consciousness is reflective consciousness of reflective consciousness}$$

it still is not really adequate as a model of most reflectively conscious experience — although it does seem to capture *something* meaningful.

In hyperset theory, one can write an equation

$$f = f(f)$$

with complete mathematical consistency. You feed f as input: $f \ldots$ and you receive as output: f. But while this sort of anti-foundational recursion may be closely associated with consciousness, this simple equation itself does not tell you much about consciousness. We do not really want to say

$$\text{ReflectiveConsciousness} = \text{ReflectiveConsciousness}(\text{ReflectiveConsciousness})$$

It is more useful to say:

$$\begin{array}{l} \text{Reflective consciousness is a hyperset, and reflective consciousness} \\ \qquad \text{is contained in its membership scope} \end{array}$$

Here by the "membership scope" of a hyperset S, what we mean is the members of S, plus the members of the members of S, etc. However, this is no longer a definition of reflective consciousness, merely a characterization. What it says is that reflective consciousness must be defined anti-foundationally as some sort of construct via which reflective consciousness builds reflective consciousness from reflective consciousness — but it does not specify exactly how.

Putting this notion together with the earlier discussion on patterns, correlations and experience, we arrive at the following working definition of reflective consciousness.

Assume the existence of some formal language with enough power to represent nested logical predicates, e.g., standard predicate calculus will suffice; let us refer to expressions in this language as "declarative content". Then we may say

Definition 4. "S is reflectively conscious of X" is defined as: The declarative content that {"S is reflectively conscious of X" correlates with "X is a pattern in S"}.

For example: Being reflectively conscious of a tree means having in one's mind declarative knowledge of the form that one's reflective consciousness of that tree is correlated with that tree being a pattern in one's overall mind-state. Figure 3 graphically depicts the above definition.

Note that this declarative knowledge does not have to be *explicitly* represented in the experiencer's mind as a well-formalized language — just as pigeons, for instance, can carry out deductive reasoning without having a formalization of the rules of Boolean or probabilistic logic in their brains. All that is required is that the conscious

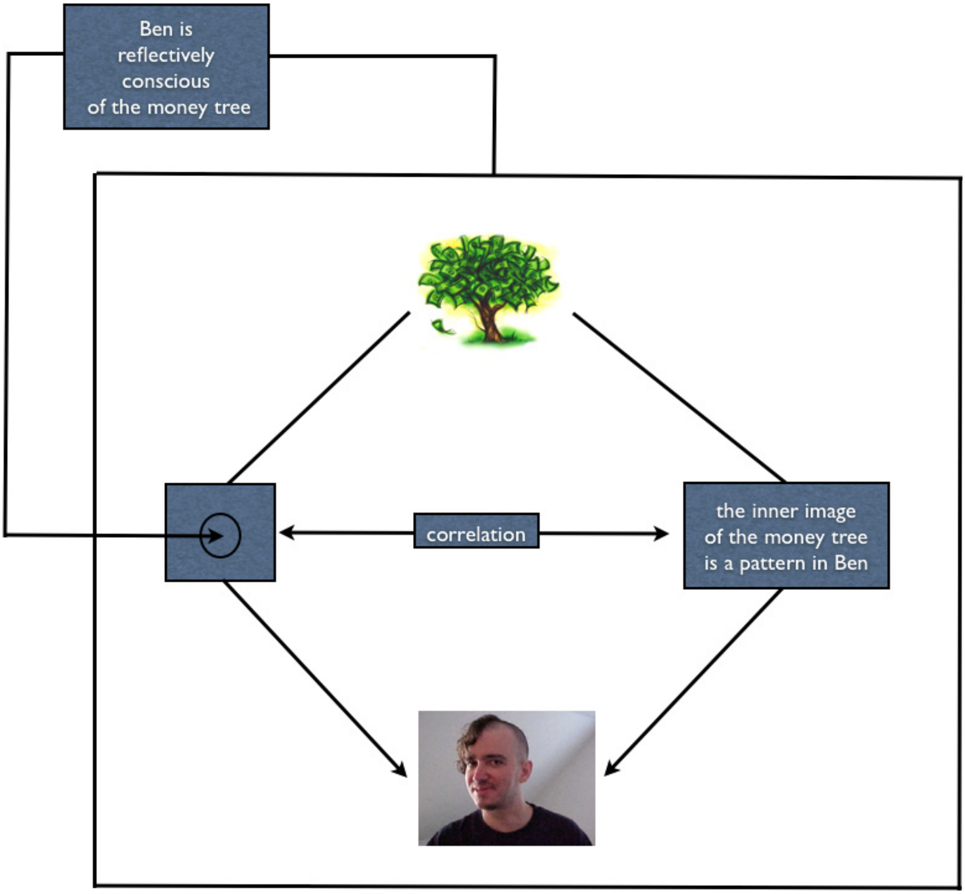
Ben is reflectively conscious of the money tree
correlation
the inner image of the money tree is a pattern in Ben

Fig. 3. Graphical depiction of "Ben is reflectively conscious of his inner image of a money tree".

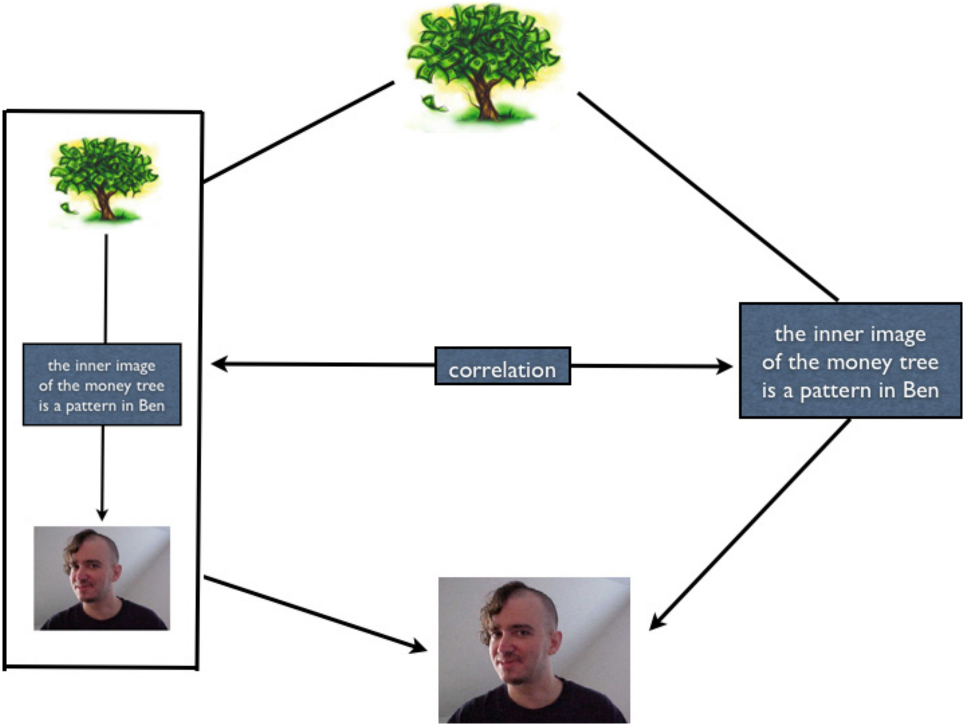

Fig. 4. First iterate of a series that converges to Ben's reflective consciousness of his inner image of a money tree.

mind has an internal "informal, possibly implicit" language capable of expressing and manipulating simple hypersets. Boolean logic is still a subpattern in the pigeon's brain even though the pigeon never explicitly applies a Boolean logic rule, and similarly the hypersets of reflective consciousness may be subpatterns in the pigeon's brain in spite of its inability to explicitly represent the underlying mathematics.

Turning next to the question of how these hyperset constructs may emerge from finite systems, Figs. 4, 5 and 6 show the first few iterates of a series of structures that would naturally be computed by a pattern containing as a subpattern Ben's reflective consciousness of his inner image of a money tree. The presence of a number of iterates in this sort of series, as patterns or subpatterns in Ben, will lead to the presence of the hyperset of "Ben's reflective consciousness of his inner image of a money tree" as a subpattern in his mind.

6. A Hyperset Model of Will

The same approach can be used to define the notion of "will", by which is meant the sort of willing process that we carry out in our minds when we subjectively feel like we are deciding to make one choice rather than another [Walter, 2001].

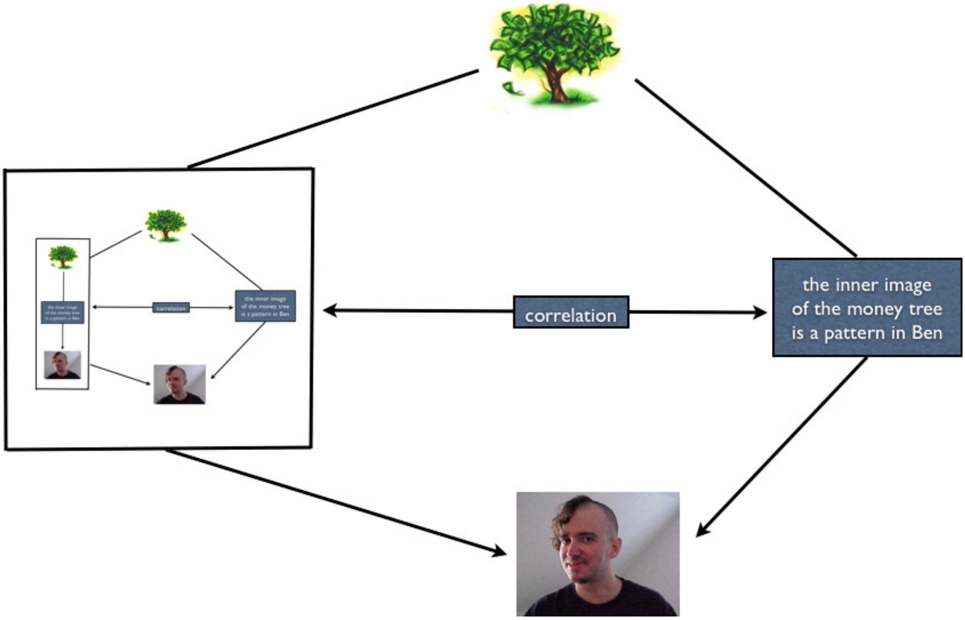

Fig. 5. Second iterate of a series that converges to Ben's reflective consciousness of his inner image of a money tree.

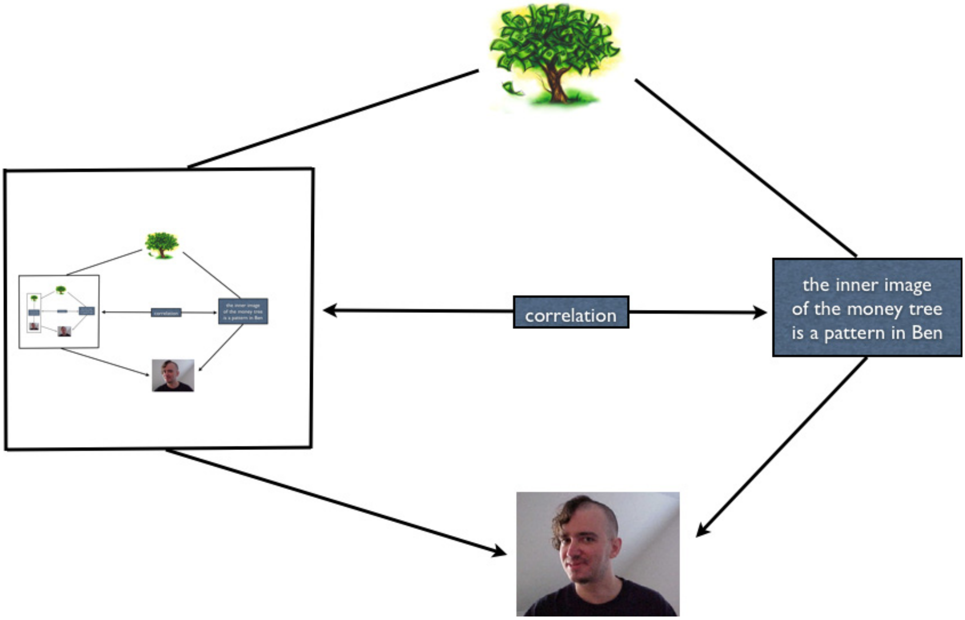

Fig. 6. Third iterate of a series that converges to Ben's reflective consciousness of his inner image of a money tree.

In brief:

Definition 5. "S wills X" is defined as: The declarative content that {"S wills X" causally implies "S does X"}.

Figure 7 graphically depicts the above definition.

To fully explicate this is slightly more complicated than in the case of reflective consciousness, due to the need to unravel what is meant by "causal implication". For the sake of the present discussion we will adopt the view of causation presented in [Goertzel *et al.*, 2008a], according to which *causal implication* may be defined as: Predictive implication combined with the existence of a plausible causal mechanism.

More precisely, if A and B are two classes of events, then A "predictively implies B" if it is probabilistically true that in a situation where A occurs, B often occurs afterwards. (Of course, this is dependent on a model of what is a "situation", which is assumed to be part of the mind assessing the predictive implication.)

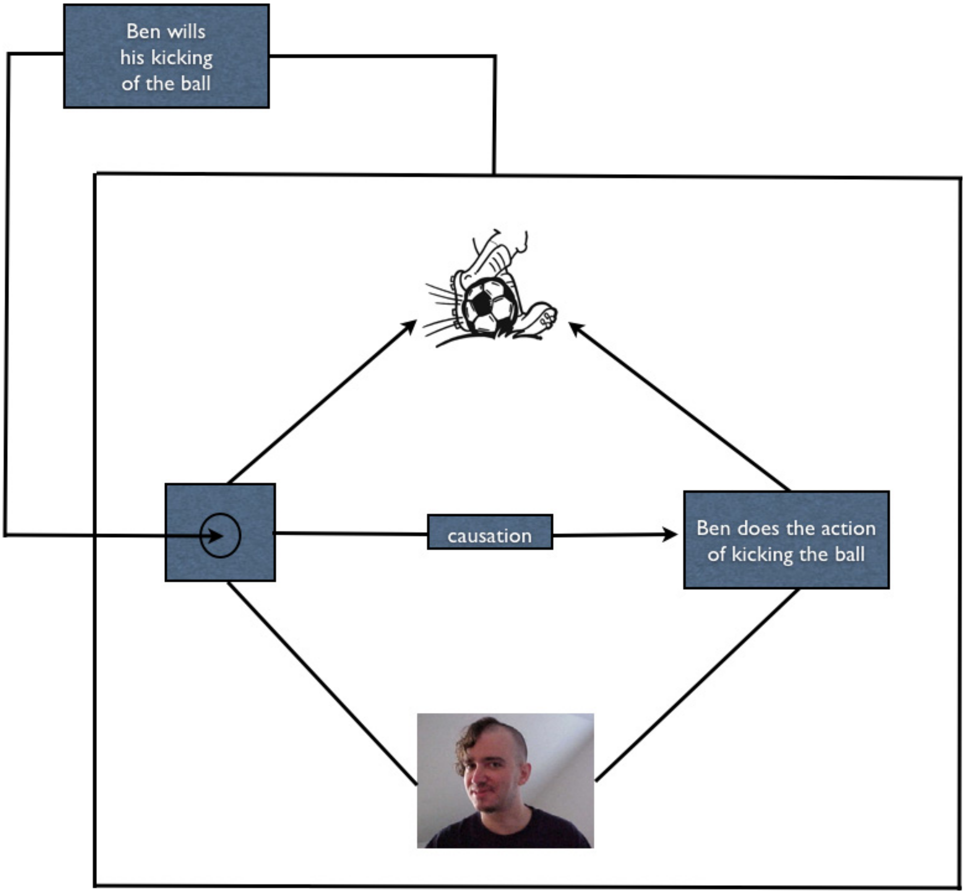
Ben wills his kicking of the ball
causation
Ben does the action of kicking the ball

Fig. 7. Graphical depiction of "Ben wills himself to kick the soccer ball".

And, a "plausible causal mechanism" associated with the assertion "A predictively implies B" means that, if one removed from one's knowledge base all specific instances of situations providing direct evidence for "A predictively implies B", then the inferred evidence for "A predictively implies B" would still be reasonably strong. (In PLN lingo, this means there is strong intensional evidence for the predictive implication, along with extensional evidence.)

If X and Y are particular events, then the probability of "X causally implies Y" may be assessed by probabilistic inference based on the classes (A, B, etc.) of events that X and Y belong to.

6.1. *In what sense is will free?*

Briefly, what does this say about the philosophical issues traditionally associated with the notion of "free will"?

It does not suggest any validity for the idea that will somehow add a magical ingredient beyond the familiar ingredients of "rules" plus "randomness". In that sense, it is not a very radical approach. It fits in with the modern understanding that free will is to a certain extent an "illusion", and that some sort of "natural autonomy" [Walter, 2001] is a more realistic notion.

However, it also suggests that "illusion" is not quite the right word. An act of will may have causal implication, according to the psychological definition of the latter, without this action of will violating the notion of deterministic/stochastic equations of the universe. The key point is that causality is itself a psychological notion (where within "psychological" I include cultural as well as individual psychology). Causality is not a physical notion; there is no branch of science that contains the notion of causation within its formal language. In the internal language of mind, acts of will have causal impacts — and this is consistent with the hypothesis that mental actions may potentially be ultimately determined via determistic/stochastic lower-level dynamics. Acts of will exist on a different level of description than these lower-level dynamics. The lower-level dynamics are part of a theory that compactly explains the behavior of cells, molecules and particles; and some aspects of complex higher-level systems like brains, bodies and societies. Will is part of a theory that compactly explains the decisions of a mind to itself.

6.2. *Connecting will and consciousness*

Connecting back to reflective consciousness, we may say that:

> In the domain of reflective conscious experiences, acts of will are experienced as causal.

This may seem a perfectly obvious assertion. What is nice is that, in the present perspective, it seems to fall out of a precise, abstract characterization of consciousness and will.

7. A Hyperset Model of Self

Finally, we posit a similar characterization for the cognitive structure called the "phenomenal self" — i.e., the psychosocial model that an organism builds of itself, to guide its interaction with the world and also its own internal choices. For a masterfully thorough treatment of this entity, see Thomas Metzinger's book *Being No One* [Metzinger, 2004]).

One way to conceptualize self is in terms of the various forms of memory comprising a human-like intelligence [Tulving and Craik, 2005], which include procedural, semantic and episodic memory.

In terms of procedural memory, an organism's phenomenal self may be viewed as a *predictive model* of the system's behavior. It need not be a wholly accurate predictive model; indeed many human selves are wildly inaccurate, and aesthetically speaking, this can be part of their charm. But it is a predictive model that the system uses to predict its behavior.

In terms of declarative memory, a phenomenal self is used for explanation — it is an *explanatory model* of the organism's behaviors. It allows the organism to carry out (more or less uncertain) inferences about what it has done and is likely to do.

In terms of episodic memory, a phenomenal self is used as the protagonist of the organism's remembered and constructed narratives. It is a fictional character, "based on a true story", simplified and sculpted to allow the organism to tell itself and others (more or less) sensible stories about what it does.

The simplest version of a hyperset model of self would be:

Definition 6. "X is part of S's phenomenal self" is defined as the declarative content that {"X is a part of S's phenomenal self" correlates with "X is a persistent pattern in S over time"}.

Figure 8 graphically depicts the above definition.

A subtler version of the definition would take into account the multiplicity of memory types:

Definition 7. "X is part of S's phenomenal self" is defined as the declarative content that {"X is a part of S's phenomenal self" correlates with "X is a persistent pattern in S's declarative, procedural and episodic memory over time"}.

One thing that is nice about this definition (in both versions) is the relationship that it applies between self and reflective consciousness. In a formula:

Self is to long-term memory as reflective consciousness is to short-term memory

According to these definitions:

- A mind's self is nothing more or less than its reflective consciousness of its persistent being.
- A mind's reflective consciousness is nothing more or less than the self of its short-term being.

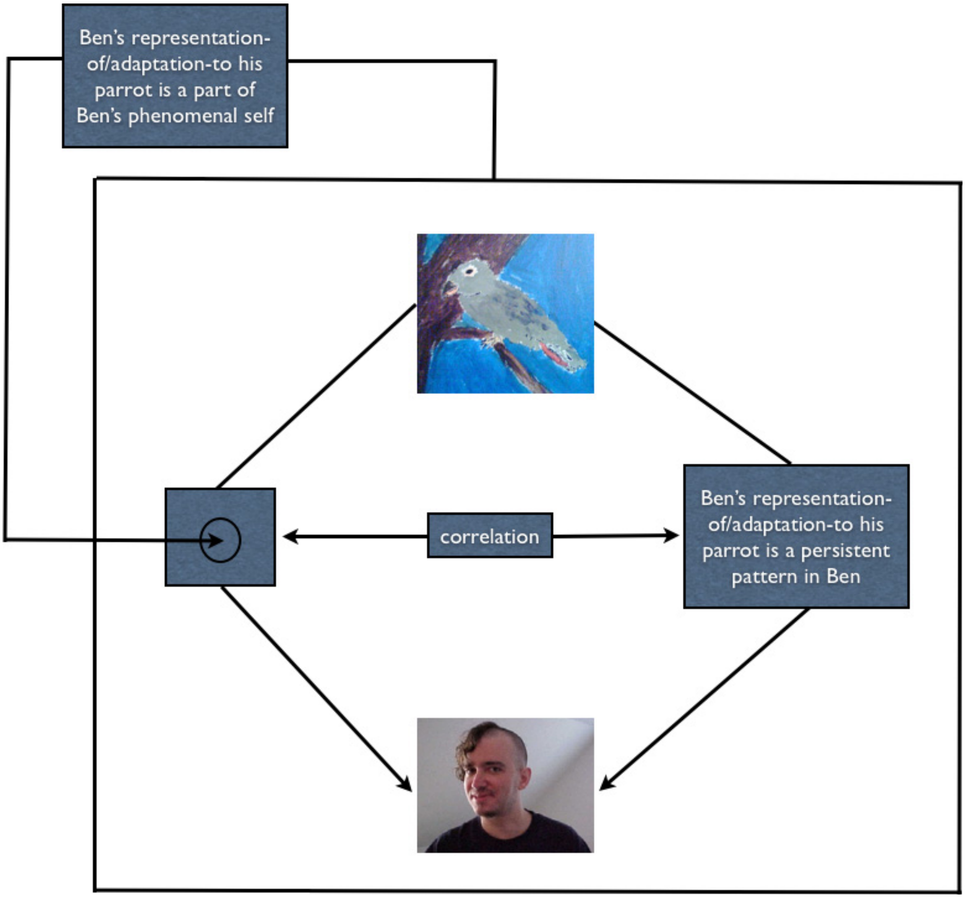
Ben's representation-of/adaptation-to his parrot is a part of Ben's phenomenal self
correlation
Ben's representation-of/adaptation-to his parrot is a persistent pattern in Ben

Fig. 8. Graphical depiction of "Ben's representation-of/adaptation-to his parrot is a part of his phenomenal self" (*Image of parrot is from a painting by Scheherazade Goertzel*).

8. Validating Hyperset Models of Experience

We have made some rather bold hypotheses here, regarding the abstract structures present in physical systems corresponding to the experiences of reflective consciousness, free will and phenomenal self. How might these hypotheses be validated or refuted?

The key is the evaluation of hypersets as subpatterns in physical systems. Taking reflective consciousness as an example, one could potentially validate whether, when a person is (or, in the materialist view, reports being) reflectively conscious of a certain apple being in front of them, the hypothetically corresponding hyperset structure is actually a subpattern in their brain structure and dynamics. We cannot carry out this kind of data analysis on brains yet, but it seems within the scope of physical science to do so.

9. Implications for Practical Work on Machine Consciousness

But what are the implications of the above ideas for *machine* consciousness in particular? One very clear implication is that digital computers probably can be just as conscious as humans can. Why the hedge "probably"? One reason is the possibility that there are some very odd, unanticipated restrictions on the patterns realizable in digital computers under the constraints of physical law. It is possible that special relativity and quantum theory, together, do not allow a digital computer to be smart enough to manifest self-reflective patterns of the complexity characteristic of human consciousness. (Special relativity means that big systems cannot think as fast as small ones; quantum theory means that systems with small enough components have to be considered quantum computers rather than classical digital computers.) This seems extremely unlikely to me, but it cannot be rated impossible at this point. And of course, even if it is true, it probably just means that machine consciousness needs to use quantum machines, or whatever other kind of machines the brain turns out to be.

Setting aside fairly remote possibilities, then, it seems that the patterns characterizing reflective consciousness, self and will can likely emerge from AI programs running on digital computers. But then, what more can be said about how these entities might emerge from the particular cognitive architectures and processes at play in the current AI field?

The answer to this question turns out to depend fairly sensitively on the particular AI architecture under consideration. Here we will briefly review the OpenCog architecture and then discuss how machine consciousness might emerge in it, according to the ideas of the previous sections.

9.1. *OpenCog and CogPrime*

CogPrime [Goertzel *et al.*, 2011] is a comprehensive architecture for cognition, language, and virtual agent control, created by the author and his collaborators during the period since 2001 (and building on their work from the 1990s). Explicitly oriented toward Artificial General Intelligence (AGI) rather than task-specific "narrow AI", and conceptually founded on the systems theory of intelligence outlined in Goertzel [2006a] and alluded to above, CogPrime is currently under development within the open-source OpenCog AI framework (see http://opencog.org and [Goertzel, 2009b]), resulting in a system sometimes referred to as OpenCogPrime or OCP. CogPrime combines multiple AI paradigms such as uncertain-logic, computational linguistics, evolutionary program learning and connectionist attention allocation in a unified cognitive-science-based architecture. Cognitive processes embodying these different paradigms interoperate together on a common neural-symbolic knowledge store called the Atomspace.

Figure 9 shows the high-level architecture of CogPrime, in the context of the OpenCogBot architecture for CogPrime-based robotics. The architecture involves the use of multiple cognitive processes associated with multiple types of memory to

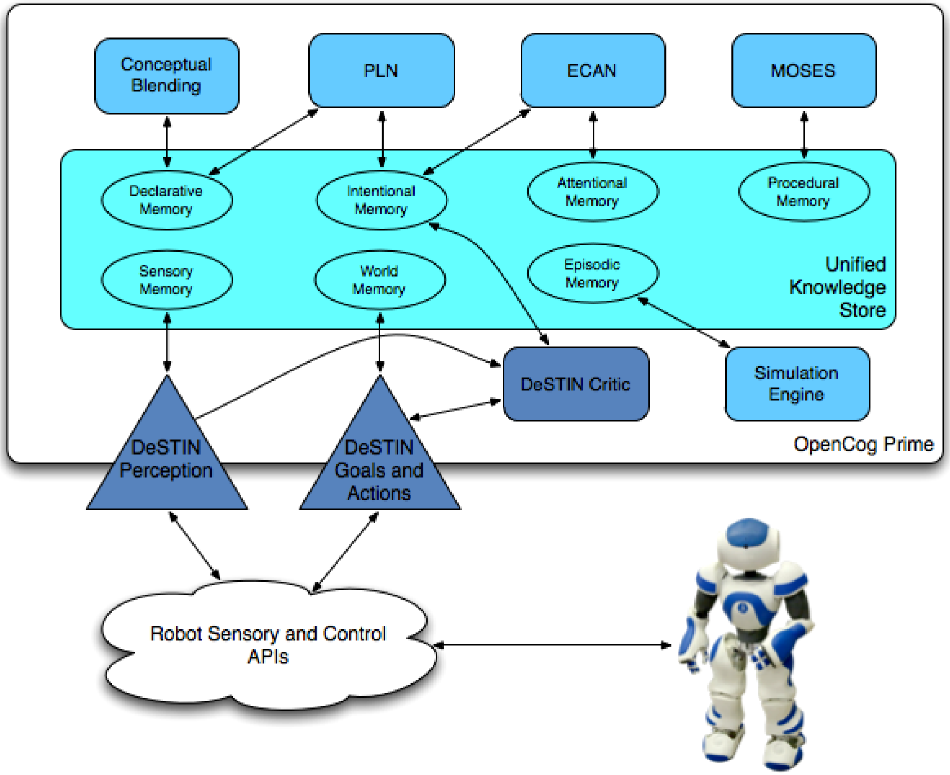

Fig. 9. Graphical depiction of "OpenCogBot", the CogPrime architecture applied to robotics.

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 are simple things such as pleasing the teacher, learning new information and skills, and protecting the robot's body.

9.1.1. *Memory types in OpenCogPrime*

CogPrime's architecture involves specialized handling of multiple memory types — the declarative, procedural, sensory, and episodic memory types that are widely discussed in cognitive neuroscience [Tulving and Craik, 2005] (and that were briefly mentioned above in the context of phenomenal self), plus attentional memory for allocating system resources generically, and intentional memory for allocating system resources in a goal-directed way. Table 1 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.).

CogPrime's declarative knowledge is called the AtomSpace, and is used as the central hub via which the various memory stores swap information. The AtomSpace

Table 1. Memory types and cognitive processes in OpenCog Prime. The third column indicates the general cognitive function that each specific cognitive process carries out, according to the patternist theory of cognition.

Memory Type	Specific Cognitive Processes	General Cognitive Functions
Declarative	Probabilistic Logic Networks (PLN) [Goertzel *et al.*, 2008a]; conceptual blending [Fauconnier and Turner, 2002]	Pattern creation
Procedural	MOSES (a novel probabilistic evolutionary program learning algorithm) [Looks and Goertzel, 2009]	Pattern creation
Episodic	Internal simulation engine [Goertzel *et al.*, 2008b]	Association, pattern creation
Attentional	Economic Attention Networks (ECAN) [Goertzel *et al.*, 2010b]	Association, credit assignment
Intentional	Probabilistic goal hierarchy refined by PLN and ECAN, structured according to MicroPsi [Bach, 2009]	Credit assignment, pattern creation
Sensory	In OpenCogBot, this will be supplied by the DeSTIN component	Association, attention allocation, pattern creation, credit assignment

is a weighted labeled hypergraph, consisting of a large network of Atoms which may be either Nodes or Links, and which have many different types. For instance, ConceptNodes represent abstract concepts or parts of networks that represent abstract concepts; SchemaNodes represent procedures; HebbianLinks represent associative relationships between other Atoms; ImplicationLinks represent logical implications; etc. There are a few dozen Node and Link types with specific semantic interpretations. Atoms are weighted with probability structures denoting uncertain truth values, and attention values indicating the ShortTermImportance (STI) and LongTermImportance (LTI) of the Atom.

In terms of the patternist perspective on mind, 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, efficiency is not a side-issue but rather the essence of real-world AGI (since as Hutter [2005] 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.)
- 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**.

As an example, when learning a skill via "trial and error" (reinforcement learning [Sutton and Barto, 1998], which is focused on procedural knowledge and often carried out in OpenCog by MOSES) proves overly slow, the system may decide to involve declarative inference and do some abstract reasoning by analogy, or may decide to invoke episodic knowledge and run mental simulations of carrying out the activity involved. It may then want to resume reinforcement learning after a bit of carrying out these other activities (or continue reinforcement learning in parallel), which implies that its procedural knowledge representation and learning methods should be able to encompass lessons learned via declarative and episodic methods.

9.1.2. *Goal-oriented dynamics in OpenCogPrime*

CogPrime's dynamics has both goal-oriented and "spontaneous" aspects; both are important for understanding how consciousness might emerge in a CogPrime system.

The spontaneous aspects are largely driven by the Economic Attention Allocation aspect of the system, according to which activity spreads around the system in complex nonlinear patterns, influenced by various perceptions, actions and cognition, but also developing and being (self-)guided by emergent transient and attractor patterns.

The basic goal-oriented dynamic of the CogPrime system, within which the various types of memory are utilized, is driven by implications known as "cognitive schematics", which take the form

$$Context \wedge Procedure \rightarrow Goal\langle p\rangle$$

(summarized $C \wedge P \rightarrow G$). Semi-formally, this implication may be interpreted to mean: "If the context C appears to hold currently, then if I enact the procedure P, I can expect to achieve the goal G with certainty p". Cognitive synergy means that the learning processes corresponding to the different types of memory actively cooperate in figuring out what procedures will achieve the system's goals in the relevant contexts within its environment.

CogPrime's cognitive schematic is significantly similar to production rules in classical architectures like SOAR [Laird *et al.*, 1987] and ACT-R [Anderson, 1996]; however, there are significant differences which are important to CogPrime's functionality. Unlike with classical production rules systems, uncertainty is core to CogPrime's knowledge representation, and each CogPrime cognitive schematic is labeled with an uncertain truth value, which is critical to its utilization by CogPrime's cognitive processes. Also, in CogPrime, cognitive schematics may be incomplete, missing one or two of the terms, which may then be filled in by various cognitive processes (generally in an uncertain way). A stronger similarity is to MicroPsi's triplets; the differences in this case are more low-level and technical and are discussed in Goertzel *et al.* [2011].

Finally, the biggest difference between CogPrimes cognitive schematics and production rules or other similar constructs, is that in CogPrime this level of knowledge representation is not the only important one. CLARION [Sun and Zhang, 2004] uses production rules for explicit knowledge representation and then uses a totally separate subsymbolic knowledge store for implicit knowledge. In CogPrime both explicit and implicit knowledge are stored in the same graph of nodes and links, with explicit knowledge stored in probabilistic logic-based nodes and links such as cognitive schematics; and implicit knowledge stored in patterns of activity among these same nodes and links, defined via the activity of the "importance" values associated with nodes and links and propagated by the ECAN attention allocation process.

The meaning of a cognitive schematic in CogPrime is hence not entirely encapsulated in its explicit logical form, but resides largely in the activity patterns that ECAN causes its activation or exploration to give rise to. And this fact is important because the synergetic interactions of system components are in large part modulated by ECAN activity. Without the real-time combination of explicit and implicit knowledge in the system's knowledge graph, the synergetic interaction of different cognitive processes would not work so smoothly, and the emergence of effective high-level hierarchical, heterarchical and self structures would be less likely.

9.1.3. *Current and prior applications of OpenCog*

OpenCog has been used for commercial applications in the area of natural language processing and data mining; for instance, see [Goertzel *et al.*, 2006b] where OpenCog's PLN reasoning and RelEx language processing are combined to do automated biological hypothesis generation based on information gathered from PubMed abstracts. Most relevantly to the present discussion, it has also been used to control virtual agents in virtual worlds [Goertzel *et al.*, 2008b], using an OpenCog variant called the OpenPetBrain (see http://CogPrime.net/example for some videos of these virtual dogs in action). These virtual dogs and (more so) their potential future elaborations are an interesting case study in machine consciousness.

While the OpenCog virtual dogs do not display intelligence closely comparable to that of real dogs (or human children), they do 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; and spontaneous exploration of their world, remembering their experiences and using them to bias future learning and linguistic interaction. These are simpler versions of capabilities we are working to demonstrate in ongoing work with OpenCog-controlled game characters and robots.

9.2. *Reflective consciousness, self and will in CogPrime*

How do our hyperset models of reflective consciousness, self and will reflect themselves in the CogPrime architecture?

There is no simple answer to these questions, as CogPrime is a complex system with multiple interacting structures and dynamics, but we will give here a broad outline.

9.2.1. *Attentional focus in CogPrime*

The key to understanding reflective consciousness in CogPrime is the ECAN (Economic Attention Networks) component, according to which each Atom in the system's memory has certain ShortTermImportance (STI) and LongTermImportance (LTI) values. These spread around the memory in a manner vaguely similar to activation spreading in a neural net, but using equations drawn from economics. The equations are specifically tuned so that, at any given time, a certain relatively small subset of Atoms will have significantly higher STI and LTI values than the rest. This set of important Atoms is called the AttentionalFocus, and represents the "moving bubble of attention" mentioned above, corresponding roughly to the Global Workspace in global workspace theory.

According to the patternist perspective, if some set of Atoms remains in the AttentionalFocus for a sustained period of time (which is what the ECAN equations are designed to encourage), then this Atom-set will be a persistent pattern in the system, hence a significant part of the system's mind and consciousness. Furthermore, the ECAN equations encourage the formation of densely connected networks of Atoms which are probabilistic attractors of ECAN dynamics, and which serve as hubs of larger, looser networks known as "maps". The relation between an attractor network in the AttentionalFocus and the other parts of corresponding maps that have lower STI, is conceptually related to the feeling humans have that the items in their focus of reflective consciousness are connected to other dimly-perceived items "on the fringes of consciousness".

The moving bubble of attention does not in itself constitute human-like "reflective consciousness", but it prepares the context for this. Even a simplistic, animal-like CogPrime system with almost no declarative understanding of itself or ability to model itself, may still have intensely conscious patterns, in the sense of having persistent networks of Atoms frequently occupying its AttentionalFocus, its global workspace.

9.3. *Maps and focused attention in CogPrime*

The relation between focused attention and distributed cognitive maps in CogPrime bears some emphasis, and is a subtle point related to CogPrime knowledge representation, which takes both explicit and implicit forms. The explicit level consists of Atoms with clearly comprehensible meanings, whereas the implicit level consists of "maps" as mentioned above — collections of Atoms that become important in a coordinated manner, analogously to cell assemblies in an attractor neural net.

Formation of small maps seems to follow from the logic of focused attention, along with hierarchical maps of a certain nature. But the argument for this is somewhat subtle, involving cognitive synergy between PLN inference and economic attention allocation.

The nature of PLN is that the effectiveness of reasoning is maximized by (among other strategies) minimizing the number of incorrect probabilistic independence assumptions. If reasoning on N nodes, the way to minimize independence assumptions is to use the full inclusion−exclusion formula to calculate interdependencies between the N nodes. This involves 2^N terms, one for each subset of the N nodes. Very rarely, in practical cases, will one have significant information about all these subsets. However, the nature of focused attention is that the system seeks to find out about as many of these subsets as possible, so as to be able to make the most accurate possible inferences, hence minimizing the use of unjustified independence assumptions. This implies that focused attention cannot hold too many items within it at one time, because if N is too big, then doing a decent sampling of the subsets of the N items is no longer realistic.

So, suppose that N items have been held within focused attention, meaning that a lot of predicates embodying combinations of N items have been constructed and evaluated and reasoned on. Then, during this extensive process of attentional focus, many of the N items will be useful in combination with each other — because of the existence of predicates joining the items. Hence, many HebbianLinks (Atoms representing statistical association relationships) will grow between the N items — causing the set of N items to form a map.

By this reasoning, focused attention in CogPrime is implicitly a map formation process — even though its immediate purpose is not map formation, but rather accurate inference (inference that minimizes independence assumptions by computing as many cross terms as is possible based on available direct and indirect evidence). Furthermore, it will encourage the formation of maps with a small number of elements in them (say, $N < 10$). However, these elements may themselves be ConceptNodes grouping other nodes together, perhaps grouping together nodes that are involved in maps. In this way, one may see the formation of hierarchical maps, formed of clusters of clusters of clusters …, where each cluster has $N < 10$ elements in it.

It is tempting to postulate that any intelligent system must display similar properties — so that focused consciousness, in general, has a strictly limited scope and causes the formation of maps that have *central cores* of roughly the same size as its scope. If this is indeed a general principle, it is an important one, because it tells you something about the general structure of concept networks associated with intelligent systems, based on the computational resource constraints of the systems. Furthermore this ties in with the architecture of the self.

9.4. *Reflective consciousness, self and will in CogPrime*

So far we have observed the formation of simple maps in OpenCogPrime systems, but we have not yet observed the emergence of the most important map: the self-map. According to the theory underlying CogPrime, however, we believe this will ensue once an OpenCogPrime-controlled virtual agent is provided with sufficiently rich experience, including diverse interactions with other agents.

The self-map is simply the network of Nodes and Links that a CogPrime system uses to predict, explain and simulate its own behavior. “Reflection” in the sense of cognitively reflecting on oneself, is modeled in CogPrime essentially as “doing PLN inference, together with other cognitive operations, in a manner heavily involving one’s self-map”.

The hyperset models of reflective consciousness and self presented above, appear in the context of CogPrime as approximative models of properties of maps that emerge in the system due to ECAN AttentionalFocus/map dynamics and its relationship with other cognitive processes such as inference. Our hypothesis is that, once a CogPrime system is exposed to the right sort of experience, it will internally evolve maps associated with reflective cognition and self, which possess an internal recursive structure that is effectively approximated using the hyperset models given above.

Will, then, emerges in CogPrime in part due to logical Atoms known as CausalImplicationLinks. A link of this sort is formed between A and B if the system finds it useful to hypothesis that “A causes B”. If A is an action that the system itself can take (a GroundedSchemaNode, in CogPrime lingo) then this means roughly that “If I chose to do A, then B would be likely to ensue”. If A is not an action the system can take, then the meaning may be interpreted similarly via abductive inference (i.e., via heuristic reasoning such as “If I *could* do A, and I did it, then B would likely ensue”).

The self-map is a distributed network phenomenon in CogPrime’s AtomSpace, but the cognitive process called MapFormation may cause specific ConceptNodes to emerge that serve as hubs for this distributed network. These Self Nodes may then get CausalImplicationLinks pointing out from them — and in a mature CogPrime system, we hypothesize, these will correlate with the system’s feeling of *willing*. The recursive structure of will emerges directly from the recursive structure of self, in this case — if the system ascribes cause to itself, then within itself there is also a model of its ascription of cause to itself (so that the causal ascription becomes part of the self that is being ascribed causal power), and so forth on multiple levels. Thus one has a finite-depth recursion that is approximatively modeled by the hyperset model of will described above.

All this goes well beyond what we have observed in the current CogPrime system (we have done some causal inference, but not yet in conjunction with self-modeling), but it follows from the CogPrime design on a theoretical level, and we will be working over the next years to bring these abstract notions into practice.

9.5. *Encouraging the recognition of self-referential structures in the AtomSpace*

Finally, we consider the possibility that a CogPrime system might explicitly model its own self and behavior using hypersets.

This is quite an interesting possibility, because, according to the same logic as map formation: if these hyperset structures are explicitly recognized when they exist, they can then be reasoned on and otherwise further refined, which may then cause them to exist more definitively … and hence to be explicitly recognized as yet more prominent patterns … etc. The same virtuous cycle via which ongoing map recognition and encapsulation leads to concept formation, might potentially also be made to occur on the level of complex self-referential structures, leading to their refinement, development and ongoing complexity.

One relatively simple way to achieve this in CogPrime would be to encode hyperset structures and operators in the set of primitives of the “Combo” language that CogPrime uses to represent procedural knowledge (a simple LISP-like language with carefully crafted hooks into the AtomSpace and some other special properties). If this were done, one could then recognize self-referential patterns in the AtomTable via standard CogPrime methods like MOSES and PLN.

This is quite possible, but it brings up a number of other deep issues that go beyond the scope of this paper. For instance, most knowledge in CogPrime is uncertain, so if one is to use hypersets in Combo, one would like to be able to use them probabilistically. The most natural way to assign truth values to hyperset structure turns is to use infinite-order probability distributions, as described in [Goertzel, 2010]. Infinite-order probability distributions are partially-ordered, and so one can compare the extent to which two different self-referential structures apply to a given body of data (e.g., an AtomTable), via comparing the infinite-order distributions that constitute their truth values. In this way, one can recognize self-referential patterns in an AtomTable, and carry out encapsulation of self-referential maps. This sounds very abstract and complicated, but the class of infinite-order distributions defined in the above-referenced papers actually have their truth values defined by simple matrix mathematics, so there is really nothing that abstruse involved in practice.

Clearly, with this subtle, currently unimplemented aspect of the CogPrime design we are veering rather far from anything the human brain could plausibly be doing in detail. This is fine, as CogPrime is not intended as a brain emulation. But yet, some meaningful connections may be drawn to neuroscience. We have discussed how probabilistic logic might emerge from the brain, and also how the brain may embody self-referential structures like the ones considered here, via (perhaps using the hippocampus) encoding whole neural nets as inputs to other neural nets. Regarding infinite-order probabilities, it is certainly the case that the brain is wired to carry out matrix manipulations, and reduced infinite-order probabilities to them, so that it is not completely outlandish to posit the brain could be doing something mathematically analogous. Thus, all in all, it seems at least *plausible* that the brain could be doing something roughly analogous to what we have described here, though the details would obviously be very different.

10. Conclusion

Now we step back from CogPrime and return to the main theme of the paper — the general modeling of reflective consciousness, self and will in terms of hypersets.

Suppose the main hypotheses presented here are validated, in the sense that these hyperset models are shown to be reasonable approximations of reflective, consciousness, self and will in artificial brains and AGI systems. Will this mean that the phenomena under discussion — free will, reflective consciousness, phenomenal self — have been "understood"?

According to our panpsychist view, the answer would seem to be "yes", at least in a broad sense — the hyperset models presented would then constitute a demonstratively accurate model of the patterns in physical systems corresponding to the particular manifestations of universal experience under discussion. And it also seems that the answer would be "yes" according to a purely materialist perspective, since in that case we would have figured out what classes of physical conditions correspond to the "experiential reports" under discussion.

The so-called "hard problem" of consciousness has been ignored here, via sticking with panpsychist or materialist views in which the "hard problem" is not an easy problem but rather a non-problem. The ideas presented here have originated within a patternist perspective, in which what is important is to identify the patterns constituting a given phenomenon; and so we have sought to identify the patterns corresponding to free will, reflective consciousness and phenomenal self. The "hard problem" then has to do with the relationships between various qualities that these patterns are hypothesized to possess (experiential versus physical) … but from the point of view of studying brains, building AI systems or conducting our everyday lives, it is generally the patterns (and their subpatterns) that matter.

Finally, if the ideas presented here are accepted as a reasonable approach, there is certainly much more work to be done. There are many different states of consciousness, many different varieties of self, many different aspects to the experience of willing, and so forth. These different particulars may be modeled using hypersets, via extending and specializing the definitions proposed above. This modeling may be done in a broad and general way, and it may also be done in a specific way, focused on specific biological organisms or specific AI systems. This suggested research program constitutes a novel variety of consciousness studies, using hypersets as a modeling language, which may be guided from a variety of directions including empirics and introspection.

Acknowledgment

I would like to warmly acknowledge Louis Kauffman for an act of kindness that occurred back in 1986, when I was a 19-year-old Ph.D. student, when he mailed me a copy of his manuscript **Sign and Space**, which contained so many wonderful ideas and drawings related to the themes considered here. Lou's manuscript was not my first introduction to the meme of consciousness and self-reference — I got into these ideas first via reading Douglas Hofstadter at age 13 in 1979, and then later via reading G. Spencer-Brown. But my brief written correspondence with Lou (this was before email was common even in universities) and his lovely hand-written and -drawn manuscript solidified my passion for these sorts of ideas, and increased my confidence that they are not only fascinating but deeply meaningful.

References

Aczel, P. [1984] "Final universes of processes," in *Proc. Math. Foundations of Programming Semantics* (Springer).

Aczel, P. [1988] *Non-Well-Founded Sets* (CSLI Press).

Anderson, J. R. [1996] *The Architecture of Cognition* (Lawrence Erlbaum Associates).

Baars, B. and Franklin, S. [2009] "Consciousness is computational: The lida model of global workspace theory," *International Journal of Machine Consciousness*.

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

Barwise, J. [1989] *The Situation in Logic* (CLSI Press).

Barwise, J. and Etchemendy, J. [1989] *The Liar: An Essay on Truth and Circularity* (Oxford University Press).

Bateson, G. [1979] *Mind and Nature: A Necessary Unity* (Penguin).

Brown, G. S. [1967] *Laws of Form* (Cognizer).

Chaitin, G. [2008] *Algorithmic Information Theory* (Cambridge University Press).

Chalmers, D. [1997] *The Conscious Mind* (Oxford University Press).

Dennett, D. [1993] *Consciousness Explained* (Penguin).

Devlin, K. [1984] *The Joy of Sets* (Springer).

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

Goertzel, B. [1994] *Chaotic Logic* (Plenum).

Goertzel, B. [2006a] *The Hidden Pattern* (Brown Walker).

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

Goertzel, B. [2009a] "Opencog prime: A cognitive synergy based architecture for embodied artificial general intelligence."

Goertzel, B. [2009b] "Opencog prime: A cognitive synergy based architecture for embodied artificial general intelligence," *ICCI 2009*, Hong Kong.

Goertzel, B. [2010] "Infinite-order probabilities and their application to modeling self-referential semantics," *Proceedings of ICAI-10*, Beijing.

Goertzel, B., Ikle, I. G. M. and Heljakka, A. [2008a] *Probabilistic Logic Networks* (Springer).

Goertzel, B. *et al.* [2008b] "An integrative methodology for teaching embodied non-linguistic agents, applied to virtual animals in second life," *Proceedings of the 1st Conference on Artificial General Intelligence* (IOS Press).

Goertzel, B., Pennachin, C. and Geisweiller, N. [2011] *Building Better Minds*, in preparation.

Goertzel, B., Pitt, J., Ikle, M., Pennachin, C. and Liu, R. [2010a] "Glocal memory: A design principle for artificial brains and minds," *Neurocomputing, Special Issue of Artificial Brain.*

Goertzel, B., Pitt, J., Ikle, M., Pennachin, C. and Liu, R. [2010b] "Glocal memory: A design principle for artificial brains and minds," *Neurocomputing, Special Issue of Artificial Brain.*

Greenfield, S. [2001] *The Private Life of the Brain* (Wiley).

Hameroff, S. [1987] *Ultimate Computing* (North-Holland).

Hameroff, S. [2010] "The conscious pilotdendritic synchrony moves through the brain to mediate consciousness," *Journal of Biological Physics.*

Hofstadter, D. [1979] *Godel, Escher, Bach: An Eternal Golden Braid* (Basic).

Hutchins, E. [1996] *Cognition in the Wild* (Bradford Books).

Hutter, M. [2005] *Universal AI* (Springer).

Huxley, A. [1990] *The Perennial Philosophy* (Perennial).

James, W. [1950] *The Principles of Psychology* (Dover).

Kauffmann, L. [n.d.] *Sign and Space.*

Krishnananda, S. [2010] "Essays in life and eternity," `http://www.swami-krishnananda.org/life/life_32.html`.

Laing, R. [1972] *Knots* (Vintage).

Laird, J., Rosenbloom, P. and Newell, A. [1987] "Soar: An architecture for general intelligence," *Artificial Intelligence* **33**.

Looks, M. and Goertzel, B. [2009] "Program representation for general intelligence," *Proceedings of AGI-09.*

Metzinger, T. [2004] *Being No One* (Bradford).

Peirce, C. [1935] *Collected Works*, Vol. 8 (Harvard University Press).

Penrose, R. [1996] *Shadows of the Mind* (Oxford University Press).

Seager, W. and Allen-Hermanson, S. [2010] "Panpsychism," in *The Stanford Encyclopedia of Philosophy*, ed. Zalta E. N.

Shear, J. and Varela, F. [2001] *The View from Within* (Imprint Academic).

Srkbina, D. [2009] *Mind that Abides: Panpsychism in the New Millennium* (John Benjamins).

Strawson, G. [2006] *Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism?* (Imprint Academic).

Sutton, R. and Barto, A. [1998] *Reinforcement Learning* (MIT Press).

Sun, R. and Zhang, X. [2004] "Top-down versus bottom-up learning in cognitive skill acquisition," *Cognitive Systems Research* **5**.

Thompson, E. [2001] *Between Ourselves* (Imprint Academic).

Tulving, E. and Craik, R. [2005] *The Oxford Handbook of Memory* (Oxford University Press).

Varela, F. [1979] *Principles of Biological Autonomy* (North-Holland).

Walter, H. [2001] *Neurophilosophy of Free Will* (MIT Press).

Whorf, B. [1964] *Language, Thought and Reality* (MIT Press).