Hyperset Models of Self, Will and Reflective Consciousness

Ben Goertzel
Novamente LLC
1405 Bernerd Place
Rockville MD 20851, USA

April 24, 2010

Abstract

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 brainlike AI system. They appear to be compatible with both panpsychist and materialist views of consciousness, and probably other views as well.

1 Introduction

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

David Chalmers [1] 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 don't 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 doesn't 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 [2]. 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* [3] 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 [4] 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*. As reviewed in [5] (or less technically in [6]), these are sets that allow circular membership structures, e.g. 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 aren't 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.

Hypersets violate the axiom of foundation and hence are not allowed in ordinary Zermelo-Frankel set theory; but they are permitted in variant set theories, for instance if one invokes the Anti-Foundation Axiom (AFA) (which, roughly speaking, permits circular membership structures that map onto graphs in a certain way). 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. All the hypersets discussed here are easily observed to be allowable under the AFA according to the the Solution Lemma stated in [5].

While the specific ideas presented here are novel, 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 [7], G. Spencer-Brown [8], Louis Kauffmann [9] and Francisco Varela [10]. 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. [10] does venture into practical psychology via porting some of R.D. Laing's psychosocial "knots" [11] into a formal non-foundational language; but this is a very specialized exercise that doesn't involve modeling general psychological structures or processes. Situation semantics [12] does analyze various commonsense concepts and relationships using hypersets; however, it doesn't address issues of subjective experience explicitly, and doesn't present formal treatments of the phenomena considered here.

1.1 Validating The Hypotheses Presented

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. Specifically:

- 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

In general, 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 isn't 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 [13] and [14]. 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.

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 [15] (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."

In the OpenCog system [16], for example, this moving bubble is achieved via economic attention network (ECAN) equations [17] 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. On the other hand, in Hameroff's recent model of the brain [18], the the brain's moving bubble of attention is achieved through dendro-dendritic connections and the emergent dendritic web.

In this 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.

3 Quantifying Pattern

To proceed further with these ideas, one must formalize the notion of "pattern." 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. The material in this section is drawn from Appendix 1 of [3]:

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*

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

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 [3].

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 [3]: 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 isn't 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*

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

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 [3] to include subpatterns as well as patterns, because this makes it simpler to describe the relationship between hypersets and minds.

3.1 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 [19] and Hameroff's [20] 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 [21]. 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

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

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.

4 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 [22] 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 cliche' that

Consciousness is consciousness of consciousness

Note that this is nicely approximated by the series

$$\begin{aligned} A \\ \text{Consciousness of } A \\ \text{Consciousness of consciousness of } A \\ ... \end{aligned}$$

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

Reflective consciousness is reflective consciousness of reflective consciousness

it still isn't 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 ... and you receive as output: f. But while this sort of anti-foundational recursion may be closely associated with consciousness, this simple equation itself doesn't tell you much about consciousness. We don't really want to say

ReflectiveConsciousness = ReflectiveConsciousness(ReflectiveConsciousness)

It's more useful to say:

Reflective consciousness is a hyperset, and reflective consciousness is contained in its membership scope

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 doesn't 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 doesn't 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 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, Figures 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.

5 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 [23].

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's meant by "causal implication." For sake of the present discussion we will adopt the view of causation presented in [**?**], 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's 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.)

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.

5.1 In What Sense Is Will Free?

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

It doesn't suggest any validity for the idea that will somehow adds a magical ingredient beyond the familiar ingredients of "rules" plus "randomness." In that sense, it's 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" [23] 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.

5.2 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's nice is that, in the present perspective, it seems to fall out of a precise, abstract characterization of consciousness and will.

6 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* [2]).

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. One thing that's nice about this definition 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.

7 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.

8 Conclusion

Suppose the hypotheses presented here are validated, in the sense proposed above. 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's 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 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.

Acknowledgement *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 PhD 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 wasn't 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

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

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

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

[4] D. Dennett, Consciousness Explained. *Penguin, 1993.*

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

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

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

[8] G. Spencer Brown, Laws Of Form. *Cognizer, 1967.*

[9] L. Kauffmann, Sign and Space. *Louis Kauffmann.*

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

[11] R. Laing, Knots. *Vintage, 1972.*

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

[13] J. Shear and F. Varela, The View From Within. *Imprint Academic, 2001.*

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

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

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

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

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

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

[20] S. Hameroff, Ultimate Computing. *North Holland, 1987.*

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

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

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

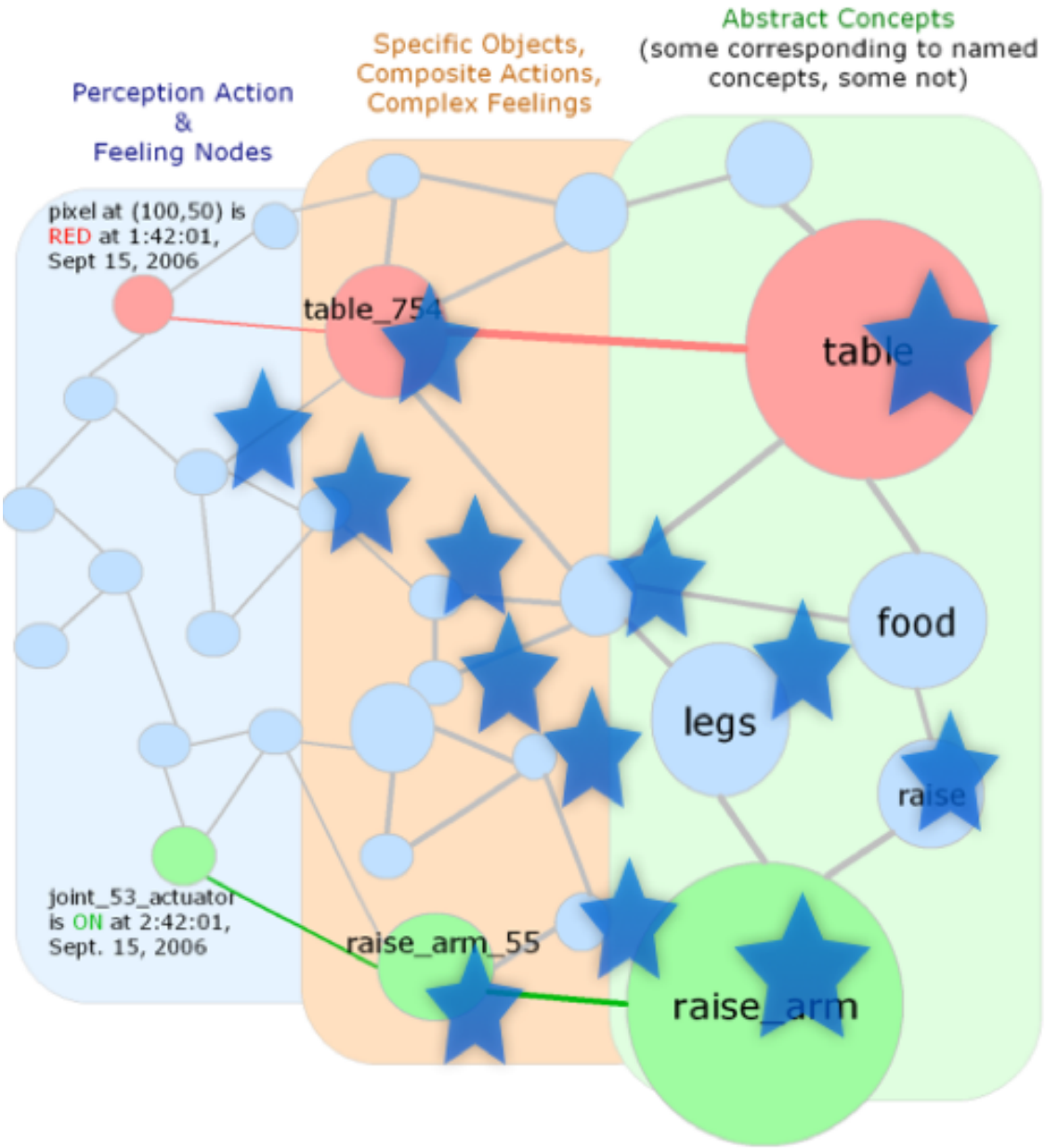

Figure 1: Graphical depiction of the momentary bubble of attention in the memory of an OpenCog AI system. Circles and lines represent nodes and links in OpenCog's memory, and stars denote those nodes with a high level of attention (represented in OpenCog by the ShortTermImportance node variable) at the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

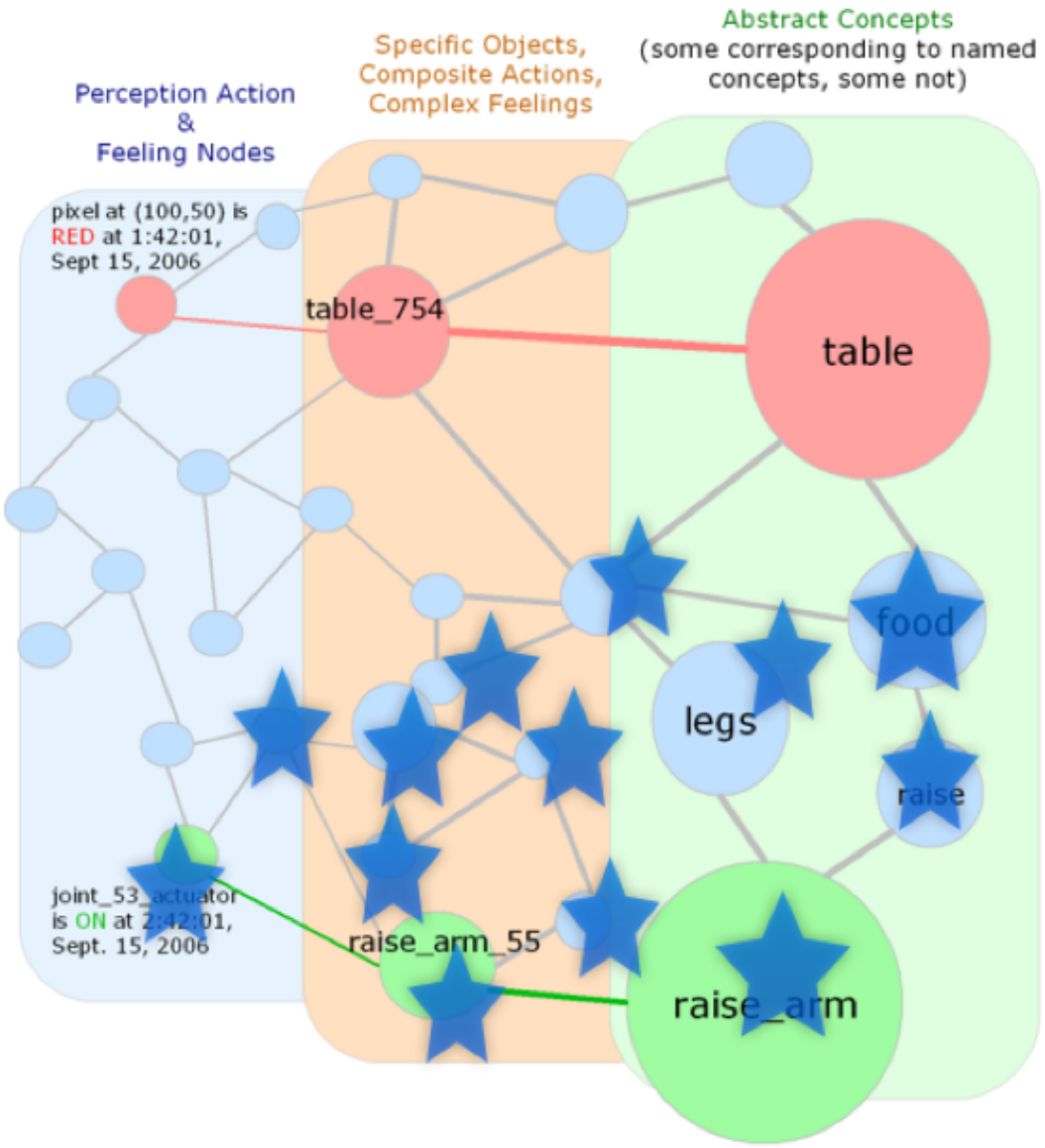

Figure 2: Graphical depiction of the momentary bubble of attention in the memory of an OpenCog AI system, a few moments after the bubble shown in Figure 1, indicating the moving of the bubble of attention. Depictive conventions are the same as in Figure 1. This shows an idealized situation where the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

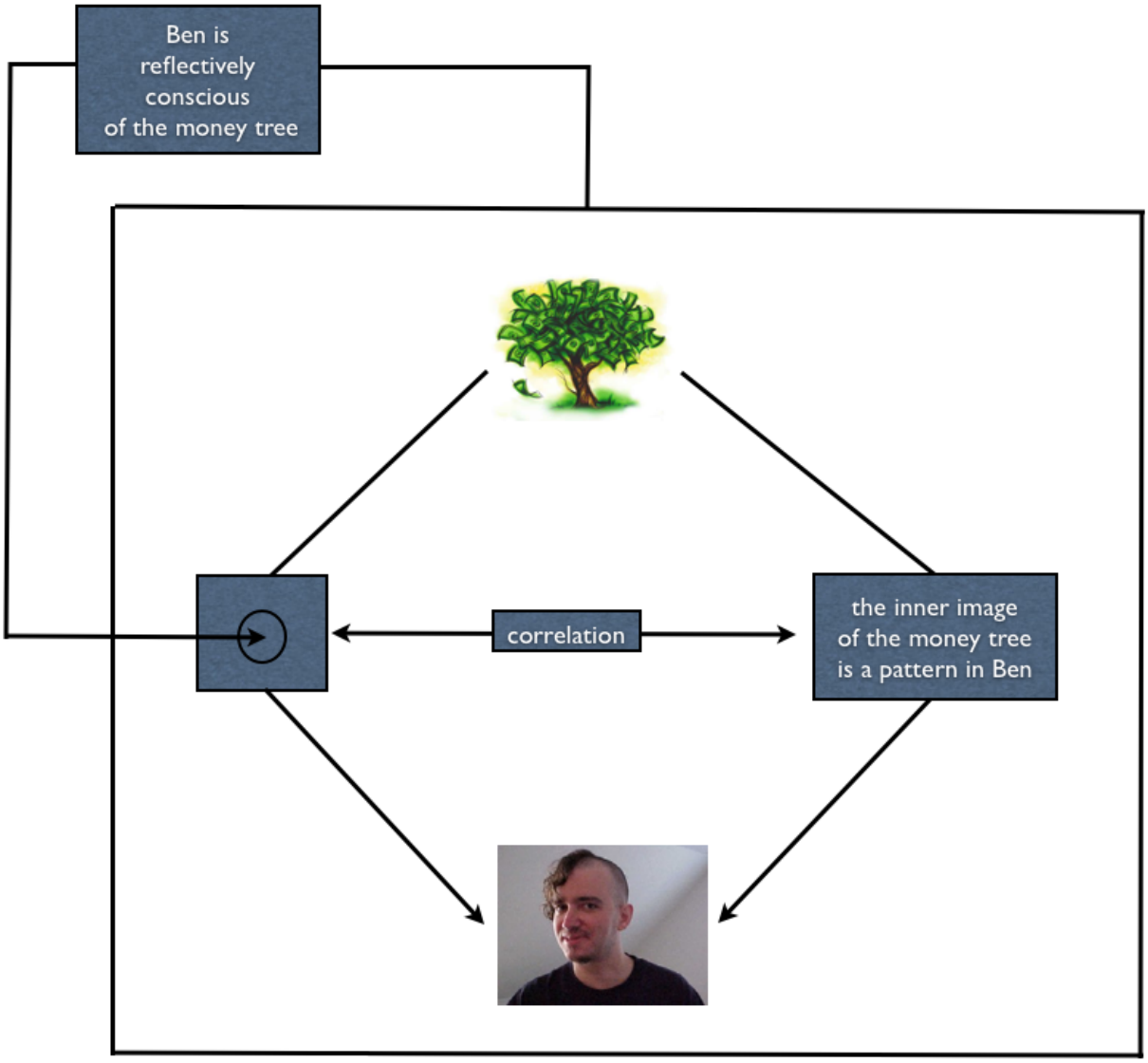

Figure 3: Graphical depiction of "Ben is reflectively conscious of his inner image of a money tree"

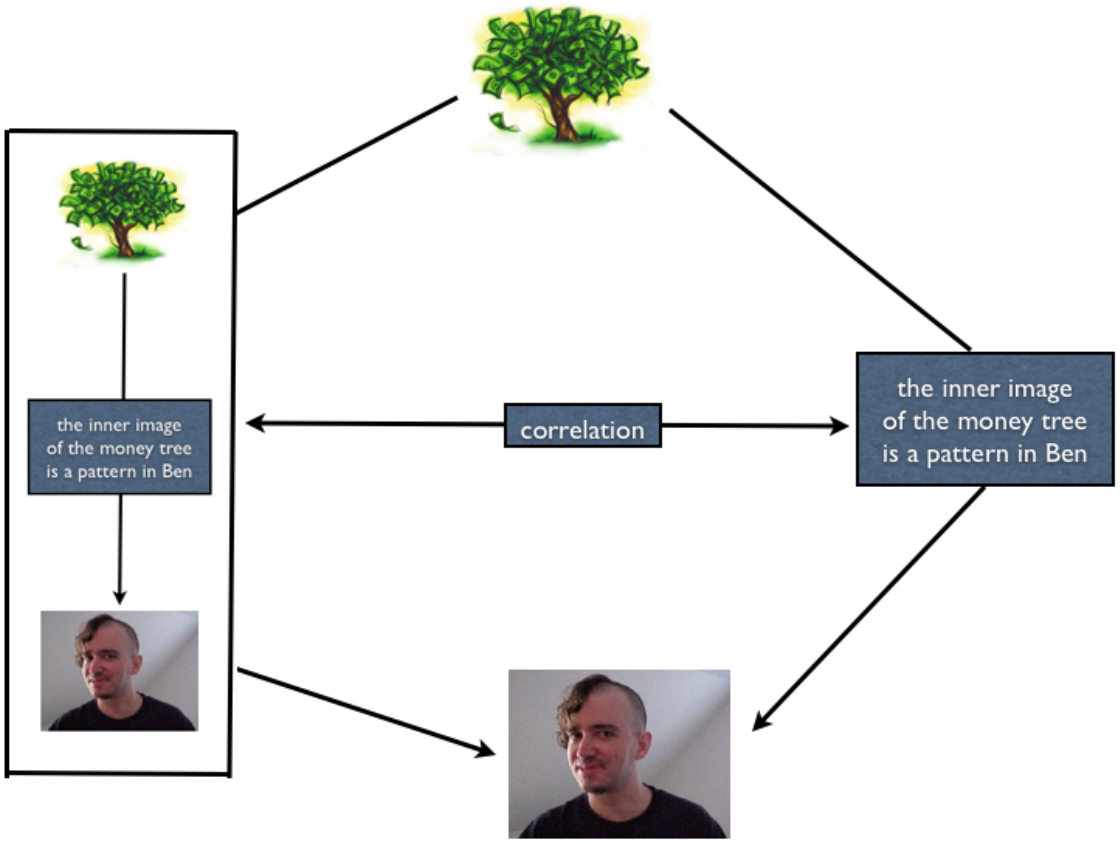

Figure 4: First iterate of a series that converges to Ben's reflective consciousness of his inner image of a money tree

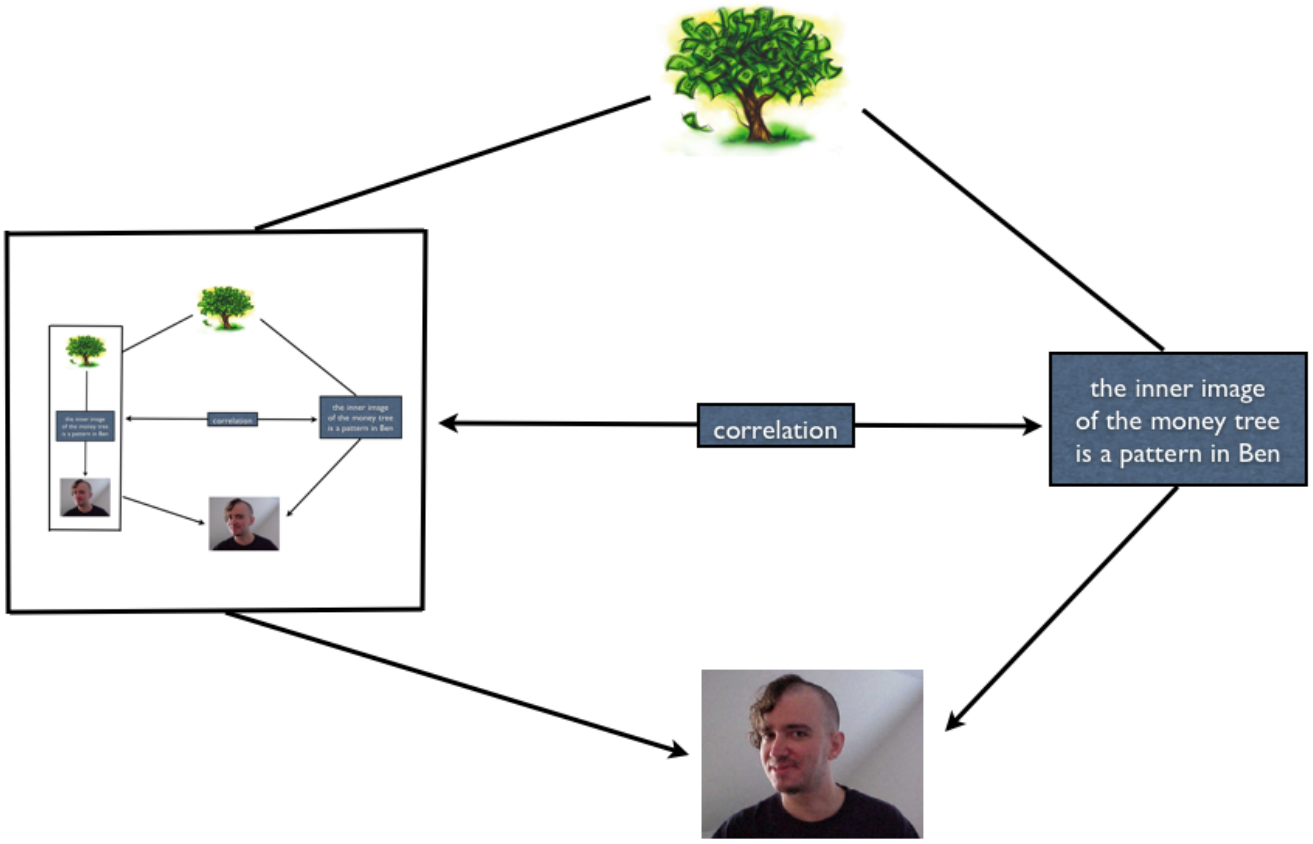

Figure 5: Second iterate of a series that converges to Ben's reflective consciousness of his inner image of a money tree

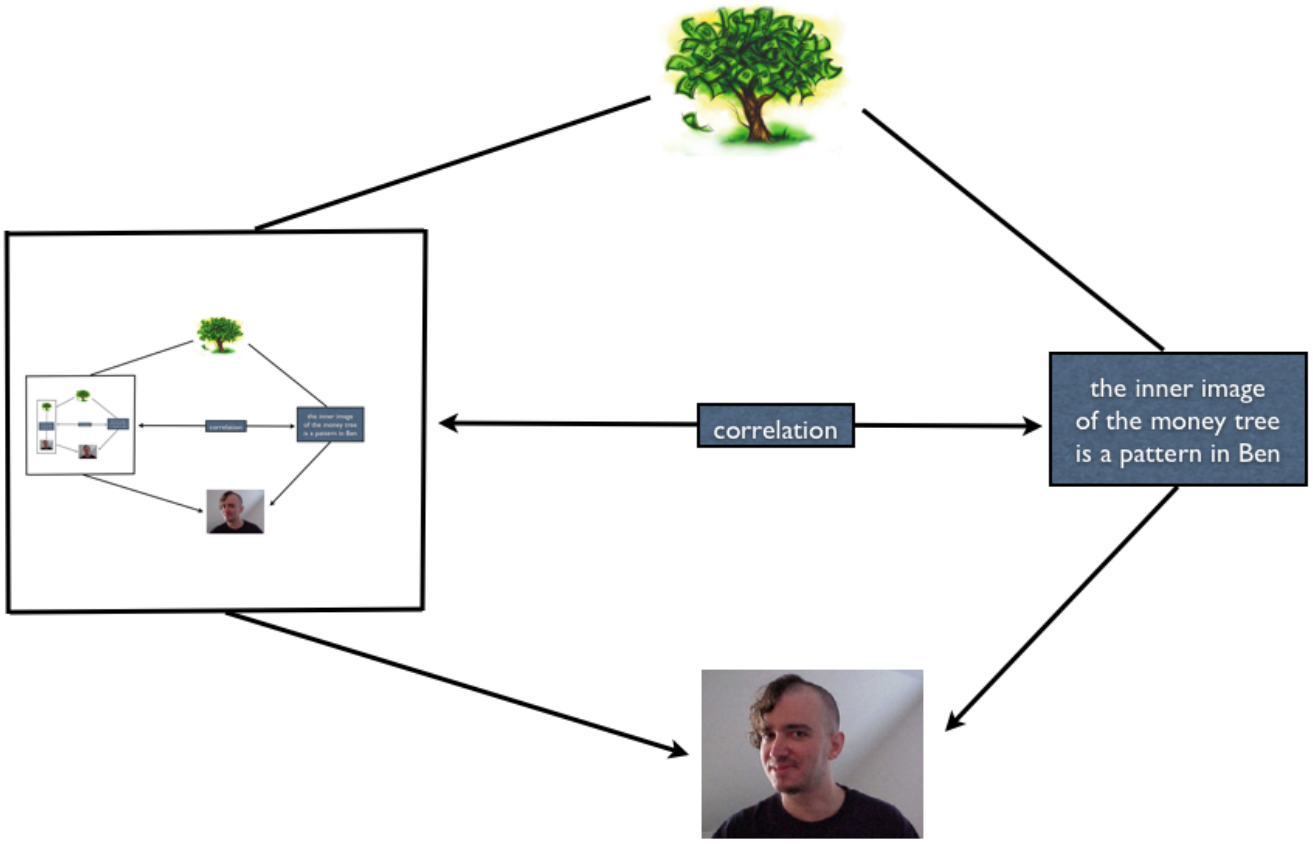

Figure 6: Third iterate of a series that converges to Ben's reflective consciousness of his inner image of a money tree

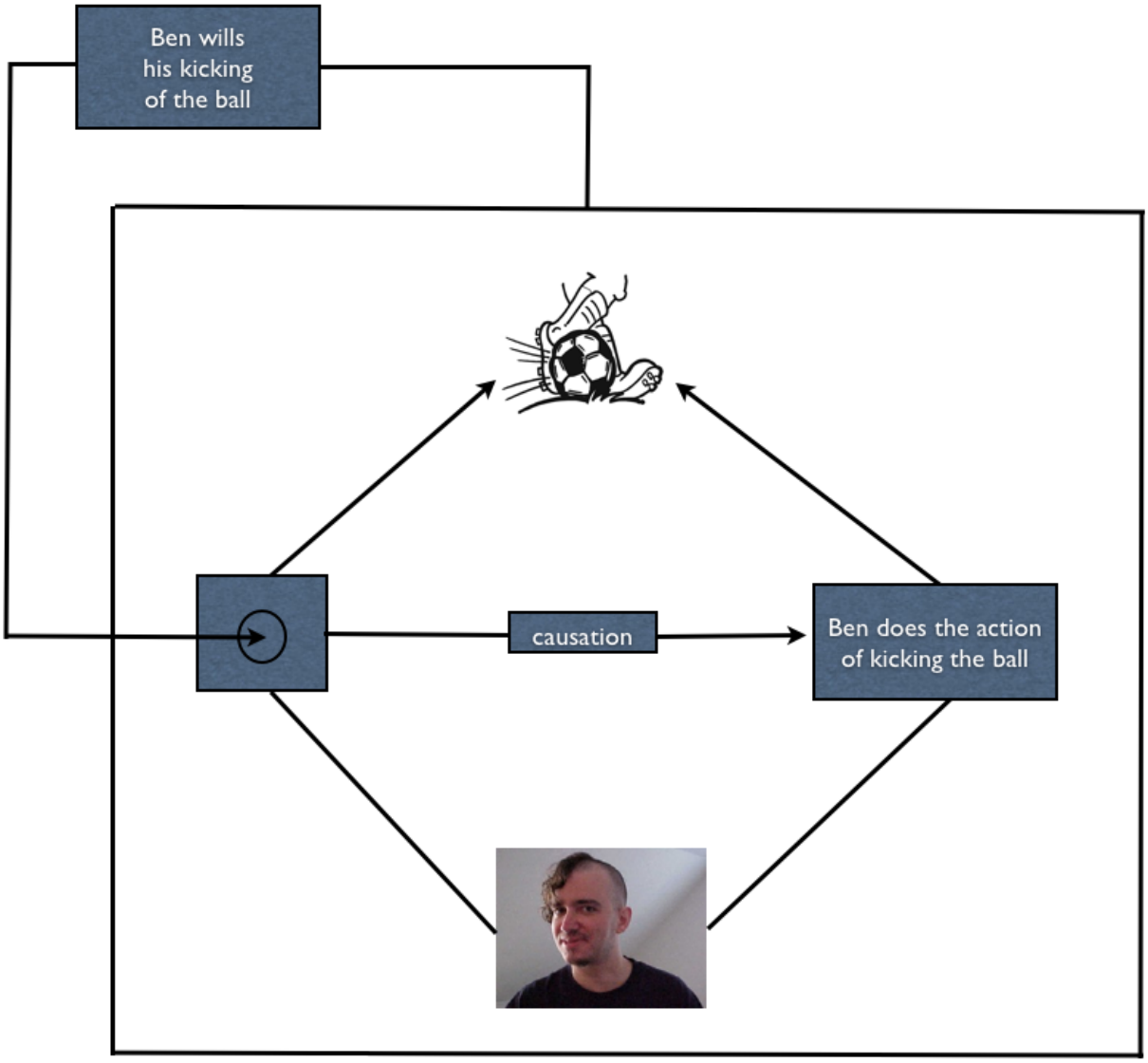

Figure 7: Graphical depiction of "Ben wills himself to kick the soccer ball"

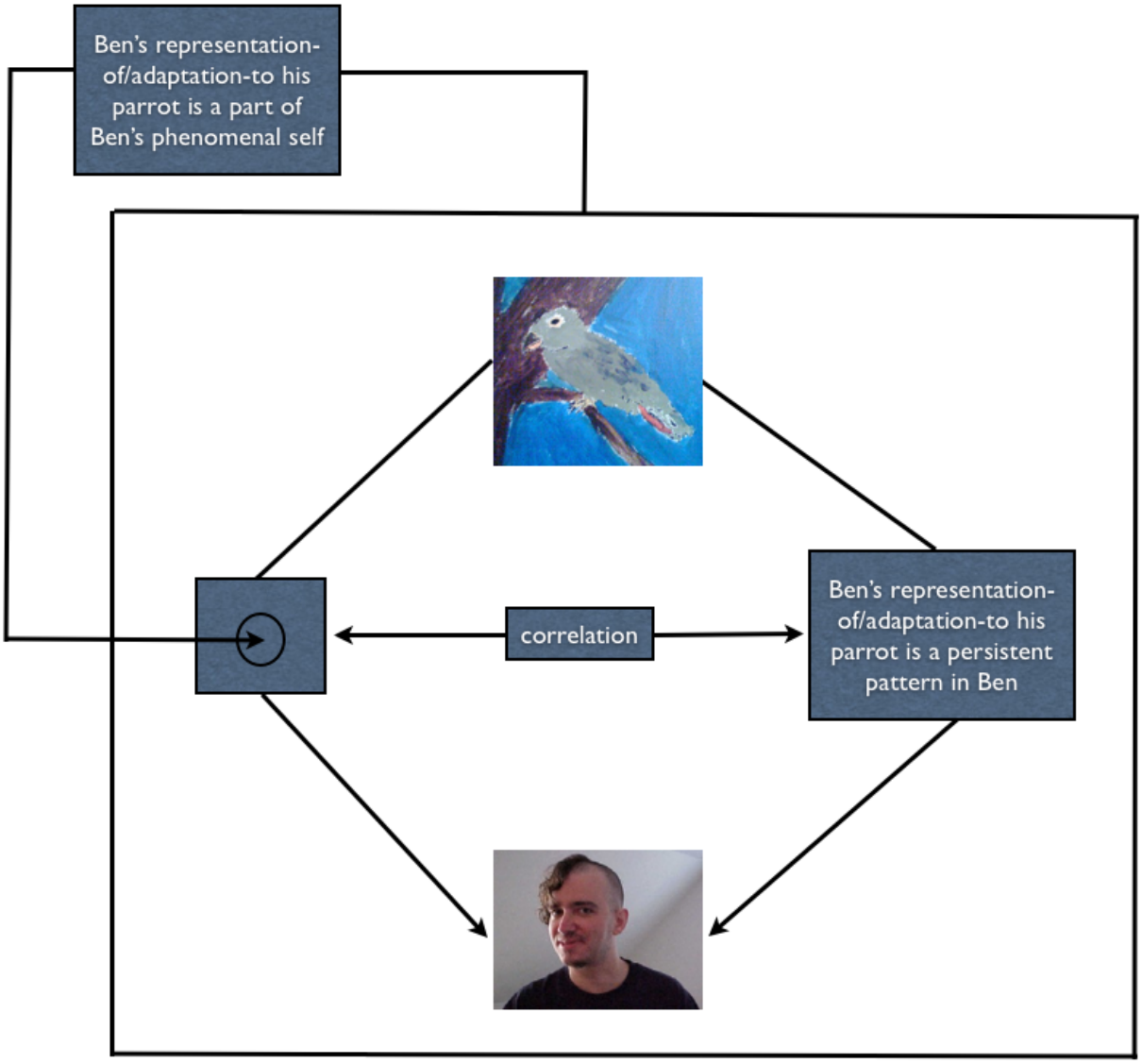

Figure 8: Graphical depiction of "Ben's representation-of/adaptation to his parrot is a part of his phenomenal self" (*Image of parrot is from a painting by Scheherazade Goertzel*)