# The Processual Self as Cognitive Unifier

Justin Brody[1], Michael T. Cox[2] and Donald Perlis[2]
[1] Goucher College, Towson MD
[2] University of Maryland Institute for Advanced Computer Studies, College Park MD

justin.brody@goucher.edu, mcox@cs.umd.edu, perlis@cs.umd.edu

**Abstract**

Cognitive Science at present is highly fragmented across multiple disciplines and even within those disciplines. It not only lacks an underlying explanatory framework but even a unifying theme or concept other than the not-very-helpful ones of mind and behavior. In this paper, we survey some ways in which a robust notion of self may be "waiting" to be fleshed out across the allied cognitive sciences in what we think may become a powerfully unifying and explanatory role.

## 1 Introduction

It is our view that progress on a number of central problems – in artificial intelligence, linguistics, logic, neuroscience, psychology, and philosophy – is stymied because of insufficient attention to the notion of the "self." Solutions to problems as disparate as the problem of relevance in AI, the binding problem in neuroscience, and the problem of determining meaning in linguistics and philosophy appear to hinge on understanding the self and its cognitive role. Specifically, *we hypothesize that the self serves as a unifier for the distinct components that make up a cognitive agent*, providing critical touchstones (e.g., of goal-directed usefulness) that – for instance – ground the distinction between relevant and irrelevant information that allows for selective attention, memory and forgetting; bind distal stimuli into mental object-constructs; support the determination of meaning and the recognition of error; and perhaps sow the seeds of consciousness. Why hasn't this issue garnered the attention we believe it deserves? Plausibly this is due to the heavily interdisciplinary nature of the concept; any investigation of self would appear to spill over numerous traditional academic boundaries and funding programs. Nevertheless, our own past work convinces us that progress on this issue can be made. What is needed is not a combination of approaches from many different areas, but rather a highly focused frontal attack on what is in essence a single cross-cutting concept that pops up in different guises in different settings. Some limited details of our approach are given below. We start by briefly sampling just six of the many far-flung issues in which a notion of self plays a key role.

1. Putnam's Theorem (Putnam, 1978) states that for any meanings that can be assigned to words, there are (many) other very different assignments that make exactly the same sentences true as under the first assignment. As a consequence, word-meaning cannot be pinned down solely by checking "facts" (i.e., which sentences are true). But then what does determine meaning, if anything at all? Perlis (1997) suggests a kind of "neurobiological" self-information may achieve just that.
2. Alan Turing (1950/1963), in initiating the study of artificial intelligence, asked what it means to understand dialog (and dismissed it as too hard, instead settling for a practical test). But if we knew more about what meaning is (as shown so puzzling in Putnam's Theorem) we would

be better-positioned to say what it is to understand the meaning of dialog (and greatly improve the performance of automated natural-language systems). Furthermore an extension of the Turing Test as proposed by Schank (1986) would be for the contestants to explain how they actually reason about the dialog. To participate in this self-explanation task (and others, see Cox, 2011), requires the first-person perspective entailed by the concept of self.

3. Perry (1979) sets out a puzzle concerning the human ability to recognize oneself as implicated in an event.[1] But what is the self, and what does it mean for such a self to recognize it as being itself? Anderson and Perlis (2005b) suggest it might be possible to operationalize a self-construct notion relevant to Perry's puzzle. Grice (1969) has emphasized a technical notion of "speaker meaning" that implicitly involves a kind of self-based concept; we suspect this is related to Perry's puzzle in subtle ways, tied to a more primitive "natural" meaning (also found in Grice) that evolved as biological systems began to form inner models. (See Kripke (2011) for more related considerations on self-acquaintance.)

4. Current automated systems behave poorly in unanticipated circumstances, not even recognizing that something is wrong (let alone that they may have a role in it). They lack any significant concept of self. If we better understood how it is that humans often manage to adapt their ideas and efforts in real time, we would have guidelines for how to design automated systems to be similarly able to recognize (and adjust or correct) their own roles when appropriate (Anderson and Perlis, 2005a).

5. Artificial systems typically cannot distinguish between relevant and irrelevant information, which slows processing and forces unnecessary storage. Many living organisms solve this problem by modulating perception and attention in task-specific, self-oriented ways (Proffitt and Linkenauger, 2013). An investigation of such self-functions in biological systems could not only greatly expand our understanding of our own behaviors and mental states (Perlis, 1997) but also contribute to progress on one of the central problems of artificial intelligence.

6. Recurrent cortico-thalamic loops have been proposed as relevant to the distinction between conscious and unconscious brain-states (Min, 2013); but little has been indicated as to how that might work. Yet if a computationally robust self-construct can be found, this might provide insight into precise neural mechanisms to seek. Indeed, the prevalence of self-oriented perception noted above suggests that a computational model of perceiving, acting, and planning that crucially incorporates a notion of self might be relevant to the neurobiology of diverse species.

This sampling illustrates self-based processes as central in the study of word-meaning, dialog understanding, agent-role recognition, error-correction, and assessing relevance. Yet such notions in the literature have cropped up only in isolated helter-skelter fashion with connections rarely noted. Evidence suggests, however, that a powerful underlying phenomenon may be involved, one that could unite many areas of cognitive science.

For example, the utterance "I was speaking French, but this sentence I am now uttering is in English." involves action execution (i.e., communication) that is explicitly (and processually) self-referential. Yet while language is important, it may also relate to the action of forming and executing a plan to hunt for food as one's energy starts to wane. The key connection is that of a process tracking its own progress in real time. While this poses real conceptual challenges, we think it is time to address them head-on.

[1] The scenario involves a shopper spotting a trail of sugar in a supermarket and circling through the store looking for the shopper responsible only to discover that it is her sugar that is leaking from her shopping cart (and thus that it is she who she seeks).

In the remaining sections, we offer some terminological clarifications about the processual self, briefly dip into the extensive but scattered literature on self-notions, and initiate a preliminary attempt to operationalize our notion of an immediate, processual self.

# 2 The Immediate Processual Self

The word "self" admits multiple and often cloudy interpretations. But we claim at the core are aspects of a process which correspond roughly to an actor and an observer. The most developed forms of self-processing will involve a real-time merging of these aspects. To foreshadow, we ultimately aim for the idea of a process that literally monitors and adjusts its own processing of the current time-slice; this we term *immediate processual self*. This, we suggest, might provide insight into Perry's puzzle: one can refer not just to one's past, but also to one's immediately present here-and-now self, the very one that is doing that referring. This is admittedly a murky notion, but one we think holds much promise.

Self-processing need not be confined to simple monitoring; processes can also change themselves. Everyday examples of this would be using your muscles to exercise themselves or using your mind to guide your own mental state toward some end (e.g., to calm yourself down). There are clear survival advantages for a process which is able to change its own functioning. By employing functioning that adapts, a process can take into account that certain things can lead toward the ability to self-regulate while others might tend to limit it. An extreme case is an intentional crash (suicide); but even that might be selected for if there are collateral survival benefits for the "self-regulation" genes.

However, our discussion so far fails to make several rather critical distinctions. To that end, we draw attention to a variety of natural/biological/computational covariances:

1. Tree rings and a tree's age;
2. An inference-engine and an automatically updated database of its activity, affording it "autobiographical" data and the abilities to pose and answer questions about that activity and to base subsequent activity on such answers;
3. A system as in 2, but in which the time-frame for the recording of activity overlaps that of the activity being recorded (the action-record is made during the action).

In item 1, while there is a strong covariance between a tree's age and the number of its rings, the tree makes no use of this "self" information. One could imagine a "smart" tree that periodically counts its rings and then decides upon reaching age 50 that it will produce fewer nuts. Such a tree would be using autobiographical data to inform its future behavior, like item 2; and as in 2, there could still be a sharp separation between monitoring (of past self) and control (of future self). So far, then, "self" does not quite catch up to the present "immediate" self, avoiding Perry's puzzle.

But when we get to item 3, the "self" that is being monitored is not entirely past and indeed could be the self-same entity/process that is doing that monitoring. Hierarchical planning in robotics has some aspects of this: partial high-level plans are made and then refined *while they are being enacted*. One way to interpret this is that still-ongoing planning is part and parcel of the same large-scale activity that is being planned.

Type-1 self is less interesting except as a possible evolutionary step toward other types of self that are important for useful cognition. Type-2 is currently more common, although much creative work remains to be done before it reaches maturity. Alternatively, type-3 is elusive; it is not even clear just what it might amount to, nor how to implement it. To summarize, the above classes of self run from the trivial to the autobiographical to the right-now self of the present moment, immediate processual self. Autobiographical self is essentially static: a running process accesses the (unchanging, as that

process runs) biographical database and reasons about it, much as so-called temporal logics reason about a static past, present, and future, with any time-passage during that reasoning being ignored. But the immediate processual self takes account of the fact that it is acting *in* time, that the present moment has a flow to it.

This "present moment" then is not the zero-duration instant of physics. It has to last long enough to allow a process to occur, i.e., a complex-enough process that it can accomplish a referring act that records itself as part of that referring. Just what this would be computationally is as yet unclear, although a later section has some suggestions. But self is a process that unfolds in time, it is not a static database or symbolic entry that exists instantaneously. Computationally, it would be like code that executes over time but monitors that same execution as it proceeds (e.g., slowing itself – including that monitoring – down to reduce the heat-generation of the computer on which it runs).

# 3 Literature Review

In general, the literature lacks anything like a unifying notion of a processual self.[2] In this section, we discuss some related conceptualizations that have appeared. Space does not allow a detailed treatment the subject deserves, although we plan a separate essay in the future. Here we merely allude to a few highlights.

Husserl (see Varela 1999) speaks of the present as holding a bit of the past and a bit of the future; an experience is experienced, it has a passage to it, a flow, even though one cannot isolate parts of it into early, middle, late. In Perlis (1997), this flow of self from being to been is used to argue for an irreducible core experience, the ur-quale, that has to it the sense of flow and little else. We conceive of it along lines similar to the "thick time" of Humphrey (2006) and Newton (2001) who emphasize the minimal time of perception, the processing that amounts to grasping an image, which has in itself a flow, an awareness of time-in-flux. We suggest that this might be nothing more than what it takes for there to be a self-processing process. Nothing more than – but that may be a lot – perhaps requiring something like self-representing and planning in real time. We have some suggestions about this in a later section below.

Thus the extended present – or thick time – is atomic, the least (physical) time in which a self-process can occur. Actually, as Newton points out, there are two kinds of time here: the physical time over which such a process occurs, and the experiential time that arises within that self-process. Possibly faster machinery and software could shorten the physical time in which such a process runs, but arguably there is a certain amount of essential processing for the formation of type-3 self, so that there might be an experiential minimum time, the thick time Humphrey refers to. Simple introspection suggests that this is less than one second, but probably at least 1/10 of a second.

We close here with a whirlwind scattershot at some other writings: Metzinger (2004) denies that an actual self can exist, as opposed to a self-appearance. But he seems to envision a static representational notion of self, not a process model as we assert. Damasio (2012) says much on aspects of brain activity that seems associated with an organism's self but little on the actual processing required. Janzen (2008, p 155) argues that subjective experience is a kind of self-awareness; we agree and here suggest ways to investigate this. See also Kriegel and Williford (2006) for a variety of views on self and consciousness. Elman (1990) describes fascinating experiments with artificial neural networks having just a tiny amount of recurrent feedback that provides a kind of memory or temporality; this significantly improves performance in surprising ways.

[2] But see Reggia (in press) for a recent related survey.

# 4 Operationalizing the Processual Self

Here we describe a processual self computationally in terms of temporality, self-referentiality, and self-modeling. We briefly define the latter terms: *self-reference*[3] refers to a process' ability to direct its actions toward itself, while *self-modeling* refers to a process' having a representation of its own functioning. Finally, we suggest how these could be implemented in active logic. These remarks will necessarily be tentative, because we are proposing a research program rather than announcing its conclusion.

Simple self-models can be created by having parameters of a process available for inspection and updating. Richer self-models can be conceived of as having a much fuller representation of a process' code. Such rich representations can be achieved via the kind of quoting mechanism available in interpreted languages (like LISP). In particular, a process can store quoted versions of its own code, modify those as needed, and apply evaluation to update itself.

The *temporality* of a processual self can be represented in temporal logics. We outline how this might be done in *active logic* (Anderson et al, 2008), a temporal logic in which reasoning takes place *in* time and can be *about* time. In particular, every deduction is marked with a timestamp, and a predicate $Now(t)$ allows for an active logic agent to access the current time in its reasoning.

We note some fundamental differences between an active logic agent and one that operates with traditional first-order logic. Most strikingly, active logic is *paraconsistent*. That is, it is possible for a sentence and its negation to both be represented in an agent's knowledge base simultaneously. Also, active logic is context sensitive not only in the sense that its knowledge base evolves over time, but also in the further sense that the application of its inference rules is sensitive to the current state of the knowledge base. For example, in first-order logic *modus ponens* asserts that if $P$ and $P \rightarrow Q$ are both in the knowledge base, then $Q$ will be inferred. Because it is paraconsistent, active logic will want this rule to be context-dependent; in particular if $P$ and $P \rightarrow Q$ are both in the knowledge base, then $Q$ will be derived only if $\neg P$ is not in the knowledge base. This context-dependence does not even seem expressible in first-order logic; the most obvious attempt would be to write $P \wedge \neg(\neg P), P \rightarrow Q \vdash Q$ but since $\neg(\neg P)$ is logically equivalent to $P$, such a rule would reduce to modus ponens.

We work in a form of active logic with a quoting mechanism (e.g., see Purang, 2001). Our goal here is to outline the mechanisms needed to create a type-3 process. For us, this is a temporal process that has a rich self-model referring to itself in a given "thick moment." We conceive of such a moment as an interval of active logic time-steps[4].

Our first task is to create a mechanism for self-reference. We accomplish this in a somewhat *ad hoc* manner by creating a label that covaries with the moment. Specifically, each moment can begin with an assertion $ThisMoment(A)$ which assigns $A$ as the label for the current moment. A process will be conceived as a sequence of such moments, and we will use a similar indexical *ThisProcess(P)* to refer to the current process. This allows any process to self-refer. For example, one could encode the notion that this very process is taking too long with a statement of the form $ThisProcess(P) \wedge$ *TooLong*($P$). One way to have such self-referring statements lead to self-modifications would be to employ global rules that govern their behavior. For example, one could have a rule of the form *TooLong*(P) ⇒ *Shutdown*($P$).

To obtain a rich self-model, the acting and observing components of a thought process should "unify". The "observer" of an active logic process corresponds to the contents of its knowledge-base at a given time while the "actions" correspond to the derivations determined by inference rules. To unify these aspects, we record both the inference rules and their applications in the knowledge base.

[3] Self reference is notoriously tricky in many respects (see Perlis, 2006) and its role in the immediate processual self may be no exception.

[4] The precise length of such a moment is unclear. In what follows we will be very loose about how long a moment can be and what kinds of activities can occur in a single moment.

The inference rules themselves thus become objects of inference, and a process has the potential to derive a theory about itself and to modify its processing as a result of observing its processing. To that end, we introduce a predicate $Applying(t, r, a, c)$ which would enter into the database if and only if at time $t$, the inference rule which quotes to $r$ was applied with the antecedents quoting to $a$ to derive the consequent quoting to $c$. This would itself be encoded as an inference rule; there are various mechanisms that could be employed to avoid an infinite regress.

As an example of how these mechanisms work, imagine that at noon an agent begins planning the rest of its day. Among its goals is a non-negotiable need to meet Mrs. Agent at 12:30 in a location 10 minutes away. During a frenzied calculation about whether to see the movie "I, Robot" at 3p.m. or 5:20p.m., the agent realizes that its planning will not end until after 12:20, and hence it cannot finish. Thinking "this planning is taking too long, I'd better abort," the planning process halts itself. Later it might draw conclusions about its own operation that would benefit it in the future. Table 1 illustrates how this might function for an active logic agent.

*Table 1: Example of Planning with a Processual Self (marginal stripes indicate the duration of individual moments; the entire table is occurring within the process* p123*)*

| **Timestamp** | **Token** | **Comment** |
|---|---|---|
| **12:00.0.0** | *ThisMoment(a0)*<br>*ThisProcess(p123)* | Begin sequence of moments dedicated to planning the rest of the day. |
| **…** | | |
| **12:00.0.12** | *ThisMoment(a1)* | A new thick moment begins. |
| **…** | | |
| **12:19.59.0** | *ThisMoment(a1674)* | At 12:19.59.0 the thick moment a1674 begins, during which the process p123 realizes that it is taking too long. |
| **12:19.59.2** | *EstimatedCompletion(12:22.0.0, p123)* | Reevaluate when the current process will end. |
| **…** | | |
| **12:19.59.5** | *Applying(*<br>*12:19.59.5,*<br>*"ThisProcess(P)∧EstimatedCompletion(t,P)*<br>*∧ t ≥ 12:20.0.0 ⇒ TooLong(P)",*<br>*"ThisProcess(p123)*<br>*∧ EstimatedCompletion(t, p123)*<br>*∧ t == 12:22.0.0",*<br>*"TooLong(p123)")* | Keep track of the inference rule being applied to conclude that this planning process will take too long. |
| **12:19.59.7** | *TooLong(p123)* | As a result, this process should be aborted. |
| **12:19.59.8** | *Shutdown(p123)* | |

We should emphasize that the above discussion and table do not begin to do justice to the underlying ideas, which will take significantly more work to flesh out. Particularly, structures and processes addressing issues of memory, monitoring, and real-time planning will all be needed, for a start. Table 1 gives only a rough hint of the flow of information-processing required. Indeed, if we are right, then this will amount to building a self, with all the attendant cognitive apparatus that goes along with it.

# 5 Conclusion

We have attempted to lay a foundation for a long-term, interdisciplinary research program. Specifically, we envision two main tasks:

A. Hypothesize the ways in which self-constructions are employed in key aspects of cognitive science, and isolate those aspects that lead to a refined version of the hypothesis.

B. Formulate tests of the hypothesis (formal, conceptual, algorithmic, and empirical), and execute those tests.

For example, one could determine a process (algorithmic or neurobiological) that can produce self-based assertions, including statements about its current activity that takes the statement-making activity into account, and then one could write a program implementing that process that *meaningfully* generates text in two languages including "I was speaking French, but this sentence I am now uttering is in English." Performing such research using the active logic approach above would certainly increase our understanding of this quite enigmatic, cognitive competence. It is our view that a scientifically-understood theory of self can play a major integrative role in the allied cognitive sciences.

# Acknowledgments

This material is based upon work supported by ONR Grants # N00014-12-1-0430 and # N00014-12-1-0172 and by ARO Grant # W911NF-12-1-0471.

# References

Anderson, M. & Perlis, D. (2005a). Logic, self-awareness and self improvement. *Journal of Logic and Computation, 15*, 22-40.

Anderson, M. and Perlis, D. (2005b). The roots of self-awareness. *Phenomenology and the Cognitive Sciences, 4*, 297-333.

Anderson, M. L., Gomaa, W., Grant, J., & Perlis, D. (2008). Active logic semantics for a single agent in a static world. *Artificial Intelligence*, *172*(8), 1045-1063.

Cox, M. T. (2011). Metareasoning, monitoring, and self-explanation. In M. T. Cox & A. Raja (Eds.) *Metareasoning: Thinking about thinking* (pp. 131-149). Cambridge, MA: MIT Press.

Damasio, A. (2012). *Self comes to mind*. New York: Vintage Books.

Elman, J. (1990). Finding structure in time. *Cognitive Science 14*, 179-211.

Grice, P. (1969). Utterer's meaning and intentions. *Philosophical Review 78*, 147-177.

Humphrey, N. (2006). *Seeing red.* Cambridge, MA: Harvard University Press.

Janzen, G. (2008). *The reflexive nature of consciousness.* Amsterdam: John Benjamins.

Kriegel, U, & Williford, K., (Eds.) (2006). *Self-representational approaches to consciousness.* Cambridge, MA: MIT Press.

Kripke, S. (2011). The first person. In S. Kripke, *Philosophical Troubles*, vol. 1, Oxford U. Press.

Min, B.-K. (2013). A thalamic reticular networking model of consciousness. *Theoretical Biology and Medical Modelling 7*(10), 1–18.

Metzinger, T. (2004). *Being no one.* Cambridge, MA: MIT Press.

Newton, N. (2001). Emergence and the uniqueness of consciousness. *Journal of Consciousness Studies*, 8, 47-59.

Perlis, D. (1997). Consciousness as self-function. *Journal of Consciousness Studies*, *4*, 509-525.

Perlis, D. (2006). Theory and application of self-reference. In T. Bolander, V. Hendricks, & S. A. Pedersen (Eds.), *Self-reference*, Stanford, CA: CSLI.

Perry, J (1979). The problem of the essential indexical. *Nous*, *13*(1), 3-21.

Proffitt, D. R., & Linkenauger, S. A. (2013). Perception viewed as a phenotypic expression. In W. Prinz, M. Beisert, & A. Herwig (Eds.), *Action Science: Foundations of an emerging discipline* (pp. 171-198). Cambridge, MA: MIT Press.

Purang, K. (2001). Alma/Carne: Implementation of a time-situated meta-reasoner. In *Tools with Artificial Intelligence, Proceedings of the 13th International Conference on* (pp. 103-110). New York: IEEE.

Putnam, H. (1978). *Meaning and the moral sciences*, London: Routledge & Kegan Paul.

Reggia, J. (in press). The rise of machine consciousness: Studying consciousness with computational models. To appear in *Neural Networks.*

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: LEA.

Turing, A. M. (1963). Computing machinery and intelligence. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 11-35). New York: McGraw Hill. (Original work published 1950)

Varela, F. (1999), The specious present: A neurophenomenology of time consciousness, In J. Petitot, F. J. Varela, B. Pacoud & J.-M. Roy (Eds.), *Naturalizing phenomenology*. Stanford, CA: Stanford University Press.

The processual self as cognitive unifier. T Cox Brody Justin. Perlis Michael.Â  We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn, and how they learn it. Specifically, we argue that these machines should (a) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned; and (c) harness compositionality and learning-to-learn to rapidly acquire and general