The Pattern Theory of Self in Artificial General Intelligence:
A Theoretical Framework for Modeling Self in Biologically Inspired Cognitive Architectures

Abstract: In an attempt to provide a unified account for a vast literature discussing a multiplicity of selves, Shaun Gallagher (2013) has proposed a pattern theory of self. Subsequent discussion on this account has led to a concern that the pattern theory, as originally presented, stands as a mere list of aspects that fails to explain how they are related in real-time. We suggest that one way to address these criticisms and further develop the pattern theory of self is by exploring how it can be used to aid research on self in artificial general intelligence, especially in the context of biologically inspired cognitive architectures. We furthermore propose a conceptual implementation for actualizing such research in regards to the LIDA (Learning Intelligent Decision Agent) cognitive model.

Keywords: Artificial General Intelligence; Self; Biologically Inspired Cognitive Architectures; Global Workspace Theory; Learning Intelligent Decision Agent (LIDA)

1. Introduction

In an attempt to provide a unified account for the multiplicity of selves discussed across a wide interdisciplinary literature, Shaun Gallagher (2013) has proposed a pattern theory of self. Subsequent discussion on this account has led to a concern that the theory, as originally presented, stands as a mere list of aspects that fails to explain how they are related in real-time. We suggest that one way to address these criticisms and further develop the pattern theory of self is by exploring how it can be used as a theoretical framework to aid research on modeling the self in artificial general intelligence. We furthermore demonstrate a realization of this kind of research in the form of a conceptual implementation of the pattern theory of self in the LIDA (Learning Intelligent Decision Agent[1]) cognitive model.

The LIDA cognitive model is a systems-level account of mind, based on Bernard Baars' Global Workspace Theory (GWT) (Baars, 1988, 2002). We understand a mind to be the control structure for an autonomous agent (AA) (Franklin, 1995, p. 412). An AA, in turn, is "a system situated in and part of an environment, which senses that environment and acts on it over time in accordance with its own agenda, so as it may affect what it senses in the future." (Franklin & Graesser, 1997). A systems-level model of minds must account for the existence of self in at least some agents. Insofar as an AA is simultaneously situated in yet distinct from their surrounding environment, we take a model of self to capture the set of capacities an AA may

[1] This is a recent name change for the LIDA model. LIDA used to stand for "Learning Intelligent Distribution Agent." The name was derived from IDA (Intelligent Distribution Agent), an earlier model that was commissioned by the US Navy to replicate the task of assigning jobs to end-of-duty sailors in a process known as distribution. The switch to Decision, as well as the addition of Learning, highlights the fact that LIDA is a general cognitive model with reach and focus beyond the distribution task of IDA.

have to maintain and perhaps understand itself as a distinct entity in the environment. These capacities may furthermore be run as processes. Thus a model of self can be broken into sub-selves that realize some or all relevant components, such as autobiographical memory, perceptual memory, self-narratives, cultural narratives, and other related phenomenon.

Research on what an implementation of self would look like is complicated by the fact that a wide variety of selves have been proposed with different theoretical and ontological commitments. The interdisciplinary literature on the topic is not only littered with many different types of self – including preconscious minimal self, no-self, embodied self, neural self, and narrative self to name a few - but many of them are logically unconnected concepts at best and contradictory at worst (See Gallagher, 2011 for collection on various sorts of self, as well as their relations).

In this paper we explore a way for better developing the self-model in LIDA that covers as many distinct senses of self as possible. In order to further develop this account, as noted above, we utilize a pattern theory as our starting point (Gallagher, 2013). According to Gallagher, "what we call self consists of a complex and sufficient pattern of certain contributories, none of which on their own is necessary or essential to any particular self" (2013, p. 3). The nature of this pattern theory of self will be explored in more detail below.

The remainder of this paper is broken into three sections. Section II offers an overview of the LIDA cognitive model and the current LIDA self-model. Section III introduces the pattern theory of self in more detail. Section IV concerns how various aspects of a "complex and sufficient pattern" of self can be used to better understand the sub-selves present in LIDA. We conclude with several considerations about future research concerning the self in LIDA.

2. LIDA Cognitive Model

LIDA is a successor to IDA (Intelligent Distribution Agent). The latter was originally commissioned by the US Navy to assign jobs for end-of-duty sailors in a process known as distribution. The subsequent work on LIDA has focused on further developing an account of how AAs can learn and respond to their environments in fluid and dynamic ways across a variety of possible tasks.

The key issue facing any AA can be summarized in the question *what do I do next*? (Franklin 1995, 412) Answering this question from an agent's point of view will be contingent on specific features of its sensory capabilities, motor actuators, and motivations. However, these differences between agents notwithstanding, any AA will also have some overlap in the general features and

mechanisms that impact how they answer this core question. For LIDA, these general features derive from GWT.

A guiding principle for GWT is continual competition for attention within the cognitive system. Attention in LIDA is in turn facilitated by both attention codelets and structure building codelets (Franklin et al., 2016). Most operations in LIDA, including those of these codelets, operate in an asynchronous manner. Notable exceptions are the actual global broadcasts and the process of action selection, which have specific triggers to begin.

A LIDA agent succeeds at the task of deciding what to do next by running through a cognitive cycle. The cycle can be understood as the core "atom" of a cognitive "moment" (Franklin, et al., 2016). It is broken into three phases: perception & understanding, attention, and action & learning. For humans, the average cognitive cycle lasts ~200-500 ms (Madl, Baars, & Franklin, 2011). While the cognitive cycle process is asynchronous, the following overview of the cognitive cycle utilizes an explanation in serial order. We make no claim that this sort of overview is meant to support modularity but only present it for the sake of a clear overview.

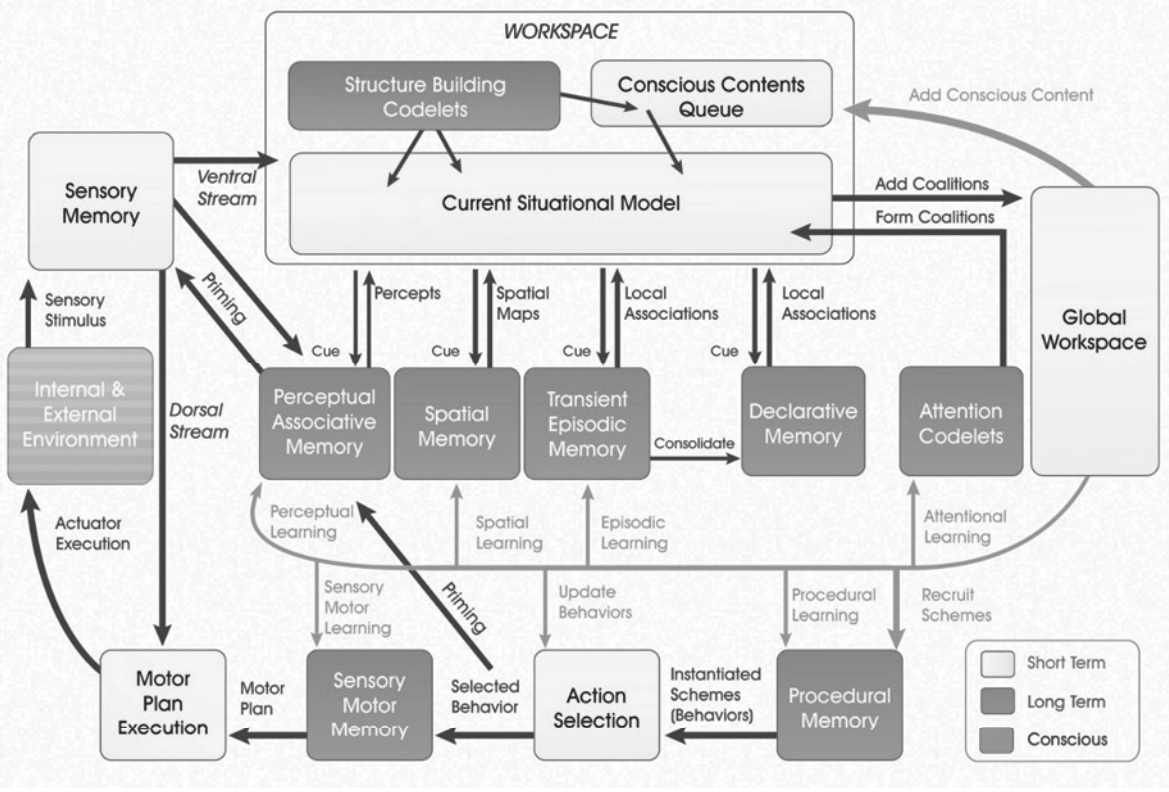


Fig. 1 The LIDA Cognitive Model

The perception and understanding phase begins with internal and/or external sensory stimulus entering into Sensory Memory. Being represented in Sensory Memory relates to updating and cuing content in the Workspace, including the Current Situational Model, and cueing Perceptual Associative Memory (PAM). The Current Situational Model, which includes any older percepts plus new inputs, will continually update itself by cued responses from PAM, Spatial Memory, Transient Episodic Memory, and Declarative Memory. Additional structuring occurs in the Workspace - which includes the Current Situational Model and the Conscious Contents Queue, with the former holding current information and the latter holding a store of recent conscious broadcasts - by structure building codelets, which survey and draw on material from the Current Situational Model and the Conscious Contents Queue.

The attention phase enables the global broadcast of the most salient information available in the Workspace. A global broadcast results from the processes of forming coalitions and the competition among these coalitions for attention. It can also be understood as the conscious phase, since a broadcast will make information functionally available throughout the system in the right sorts of ways to be considered functionally conscious (Franklin, 2003).

Not all information in the Workspace is necessarily or automatically broadcast. Attention codelets instead survey the Current Situational Model with an eye towards salient features or structures that they will attempt to bring to consciousness. These structures are built into a various coalitions by attention building codelets. The winning coalition is then broadcast globally, making information available to a variety of other subsystems.

The action and learning phase serves the dual purposes of allowing the various modules in LIDA to learn, as well as for Procedural Memory, the Action Selection module, and Sensory Motor Memory to consider, select, and enact relevant behaviors for action in light of the situation at hand. Learning occurs insofar as each module draws information from the broadcast relevant to updating its underlying data structures.

With this background of the cognitive cycle in place, we now turn to how a self fits into LIDA. Some notion of self already plays a role in asking what do *I* do next? In order to keep from begging the question or assuming a role for self at the outset, however, we can rephrase this question in terms of "What to do next?" In order for the question to make sense, especially to the agent asking it, there has to be some sort of self/non-self distinction already at play.[2]

[2] Two additional points must be flagged here. First, this view is compatible with the idea that empathy and intersubjectivity may be core components for the development of a self, especially in the case of humans. Second, we hold that it is clearly possible to cultivate certain habits, capabilities, and practices that erase an individual's awareness of this self/non-self boundary, particularly at the level of conscious experience. There are also cases

Previous work on the self and LIDA has focused on the connection between consciousness and self as well as the implementation of a self-system (Ramamurthy & Franklin, 2011; Ramamurthy, Franklin, & Agrawal, 2012). Since LIDA is an extension of GWT, a LIDA agent will necessarily be functionally conscious. This sort of consciousness is possible without the agent being able to experience any access of the contents of consciousness; in other terms, some components of the self-model may be subpersonal and/or preconscious.[3]

Ramamurthy, Franklin, and Agrawal (2012) break the self into three general components: Proto-Self, Minimal (core) Self, and Extended Self. The latter two of these have additional breakdowns, as can be seen in Figure 2.

The proto-self is drawn directly from the work of Anthony Damasio (1999). According to Damasio, the proto-self relates to specific neural patterns of activity that represent the immediate physical state of the organism. Information relevant to the proto-self includes neural and hormonal signals from bodily and visceral changes.

The minimal (core) self concerns a basic sense of self/non-self. It was further broken down into self-as-subject, self-as-experiencer, and self-as-agent. These three terms can alternatively be understood as the acting self, the experiencing self, and the self that can be acted on by other entities in the environment, respectively. The implementation of these three senses of self is heavily tied into PAM, especially insofar as one of its event nodes has various attributes related to the agent, subject, and action relevant for a given event.

All three subselves within the minimal self can be experiencing selves. However, while all three types of subselves are part of the minimal self, we consider the self-as-experiencer as a sort of default within the minimal self when the other two are not in action. Sometimes this part of the minimal self may exist within the structure of PAM without being globally broadcast, in which case it may not come into consciousness, which is different from any inherent qualities to the experiential mode of the self-as-experiencer. Thus there remains a self-as-experiencer when one is sitting quietly and relaxing, even if there is no self that is acting or directly being acted upon.

The extended self includes the autobiographical self, the self-concept, the volitional (executive) self, and the narrative self. The autobiographical self develops directly in relation to long term

---

where a self may no longer be present to, e.g, practiced meditators. Neither GWT in general nor the LIDA model in particular require a metaphysically robust notion of self as substance or thing to incorporate a self-model.

[3] Subpersonal can refer to processes that take place below conscious awareness. Preconsciousness has a more technical definition for the LIDA model and is both defined and explained more in Section 3.2

episodic memories about oneself. The self-concept consists of self beliefs and intentions, especially in relation to personal identity. The volitional self deals with executive functions and control. Finally, the narrative self is about the ability of an agent to report their actions, intentions, and related mental features, regardless of how any questions about the accuracy of those report. Several of these selves draw in part on linguistic skills.

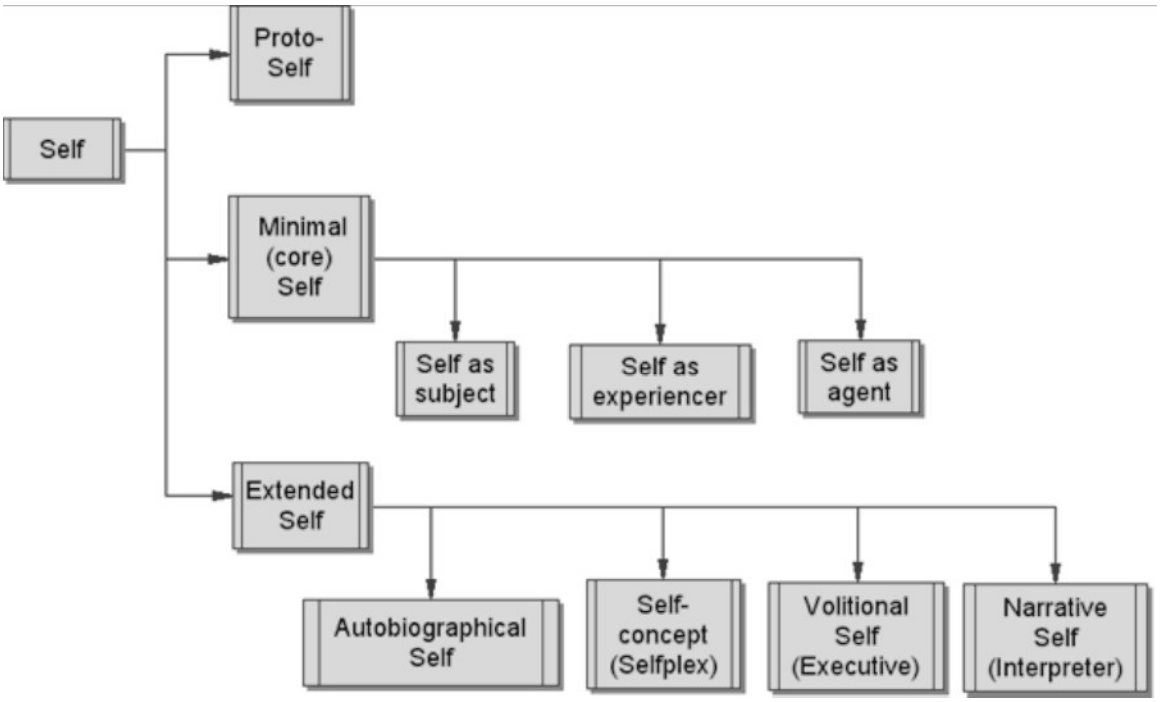


Fig. 2 The LIDA Self-Model

3. Pattern Theory of Self

Gallagher's (2013) pattern theory of self can be understood in relation to the pattern theory of emotion (Newen, Welpinghus, & Juckel, 2015) and has been applied and critiqued in several other recent contexts (Beni, 2016; De Haan, Rietveld, Stokhof, & Denys, 2017; Dings & de Bruin, 2016; Kyselo, 2014). The sense of pattern here is not a geometrical pattern. Instead, it is the notion that self is a cluster concept or family resemblance (Wittgenstein, 1953). A cluster concept is constituted by a system of aspects that lacks any strictly necessary conditions (Sloman, 2002). In other words, a cluster concept is made up of several jointly sufficient conditions from which any instance of the phenomenon will be realized. Some of these jointly sufficient aspects are more often bundled together. Likewise, some token instances will have most or all of the aspects while others will have few or perhaps only one.

Games are a classic example of a cluster concept. What are the core aspects of games? Some main candidates include an unchanging set of rules, an element of competition between two or

more players, winners-losers, and a time limit. Yet considering the wide variety of games – football, baseball, chess, trivia, Tetris, Leader/Follower, to name but a few – it is clear that not all games will share many of the same aspects. Rules between games vary widely. Some games, such as Solitaire, don't have any clear sense of competition, especially if we take competition to be between two or more agents. While many professional sports games have strict time limits, there are many more types of games that can go on indefinitely. Some games, like Celebrity Guess-Who, also don't have any clear winners or losers.

As noted in the introduction, the literature is full of different senses of what one could mean by self. For example, we can consider the differences between the proto-self and the narrative self. The proto-self is solely concerned with a pattern of relevant neural patterns of activity (Damasio, 1999). The locus of interest is thus with the activity that occurs in the brain. In contrast, the narrative self concerns a wider swath of phenomenon, such as the stories one has about what makes them who they are. The features of the narrative self can be realized as memories contained within the agent, but can also include external features such as stories they are told by friends and family, photographs they carry around, and diaries or notes that they keep. Sometimes the active manipulation of these external features are best thought of as shaped by the narrative self, while other times they may be thought of as part of the narrative self insofar as things like music, writing, and even other people can act as extended bases for affects and memories (Cochrane, 2008; Colombetti & Roberts, 2015).

As this brief example suggest, there are many different ways we can understand the patterns involved with self across different instances of selves. Explaining their exact details requires engaging with different levels of analysis. These levels, in turn, track differences within a specific type of self, among different types of selves, and what links these different types of selves together under the concept of self in the first place. Following Gallagher (2013), we consider our explanation in terms of three different levels of the pattern theory of self: the token level, the type level, and the meta-theoretical level.

Starting at the token level, any specific instance of a self or sub-self will be made from different components. For example, we have good reasons to believe that several different people each have an autobiographical self. However, there are different ways to realize and maintain such autobiographical selves. Some components that are important for one are non-existent for another. While the pattern of an autobiographical self is a cluster of jointly sufficient components, fixing what counts as an individual autobiographical self is a different project from understanding what makes an autobiographical self a distinct kind of self in the first place. When we start to explore what makes autobiographical selves tokens of a specific type, we move from a token level of analysis into a type level of analysis. At the type level we trace out features that make different types of selves distinct from each other. For example, the features of an autobiographical self will not be completely identical to those of a narrative self or a proto-self, over and above the differences in how distinct autobiographical selves are realized.

Asking what features of self make it a specific type of self also relates to the question of what a self is in the first place. Asking this question in turn involves a shift to the third level of analysis,

which Gallagher (2013) calls the meta-theoretical level. It is meta-theoretical because the analysis remains agnostic about the existence of any particular type of self. Instead, it is concerned with what aspects will capture as many different notions of self as possible. While the proto-self will not be completely co-extensive with the autobiographical self, there should be good reasons for us to consider them both kinds of self instead of something else, for instance.

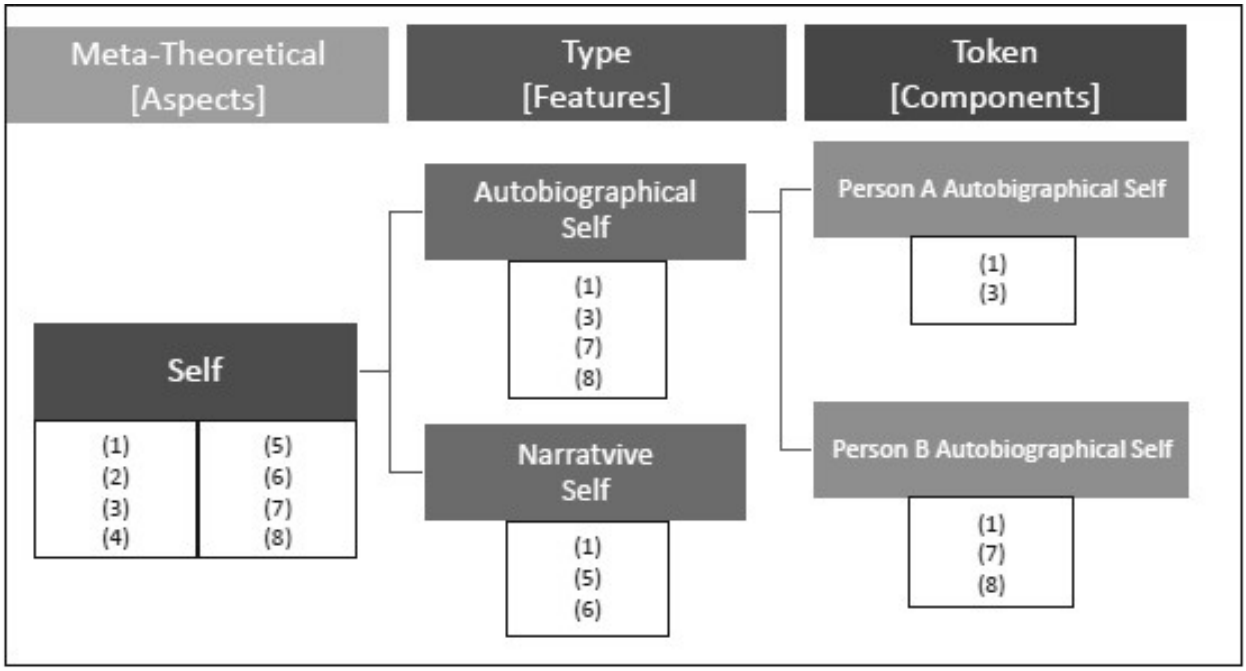


Fig 3. The Three Levels of the Pattern Theory of Self

This figure shows how the three different levels of the pattern theory of self are related. Each type *of self will have specific features that are drawn from the possible aspects of self as specified at the* meta-theoretical *level. At the same time, since the relevant patterns are jointly sufficient clusters of properties, it is possible that different* token *selves of the same type will have various forms of realization. Important details of the specific dynamics and conceptual implementation of these aspects, features, and components are not captured here but will be explored in more detail below (see section 4 in particular).*

Developing a self-model in cognitive architectures directly engages all three levels of analysis. For instance, in LIDA, the construction of LIDA agents requires careful consideration about the token level of selves and the types of self the specific agent will or will not have. Likewise, the LIDA computational framework and conceptual model, which are both broad enough to capture minds for AAs of various kinds, need to capture the variety of different selves such agents may have, which involves both the type level as well as the meta-theoretical level. The self-model introduced briefly in the previous section flexibly captures multiple elements often associated with the self and introduces how these various senses of self can be implemented in a LIDA agent. However, several questions remain. For instance: is there a useful distinction to be made between minimal (core) self and self-as-experiencer? Where is phenomenal consciousness in this self-model? How strong is the distinction between autobiographical self and narrative self,

especially since both processes include a certain interpretation of the agent's past life in relation to their current view of self? Addressing these questions requires taking stock first and foremost about the pattern of self at the meta-theoretical level, but will involve the other two levels to some degree as well

We take Gallagher's (2013) meta-theoretical list of aspects for the pattern theory of self as our starting point without additional analysis of their necessity, nor proposing any additions or subtractions to the list. LIDA, in turn, provides a useful sandbox of processes relevant to minds in this attempt to ground the variety of selves within one model. Showing how these different aspects can be implemented in LIDA will be important for the variety of selves that may be relevant for different agents. For the remainder of this section, we introduce and elaborate on these aspects, occasionally referring to the structure of LIDA to help better understand the self-model reconsidered in section 4.

3.1 Minimal Embodied Aspects

These include the biological and ecological features of a system that make it possible for an agent to distinguish between itself and what is not itself. For instance, an egocentric frame of reference will allow a system to take a first-person perspective and navigate peripersonal[4] space.

Some features of minimal embodiment will not be explicitly coded in the LIDA model or LIDA framework, since they will be contingent on specific features of the agent's body. For instance, robust proprioceptive feedback may be an important part of minimal embodied aspects for one organism (e.g. a human animal) while not for another (e.g. a software agent). Yet there are at least two important components already present in the LIDA model that will cover the main features of minimal embodied aspects: a self-node in Perceptual Associative Memory (PAM) and a difference between allocentric and egocentric spatial representations (citations below).

The LIDA model has a self-node based in PAM. Following the current node-link structure of event representation in LIDA (Franklin, et al., 2016; McCall, Franklin, & Friedlander, 2010), an event representation can have agent, subject, action, location and other kinds of links to an event to specify those roles as they relate to the event in question. The self-node can be incorporated in various structures, such as specified by these agent and/or subject links. This along with the location link can assist in understanding the location of self with respect to a frame of reference. Since this aspect is primarily concerned with a basic, functional difference between self and non-

[4]Peripersonal space refers to the region of space within physical reach of an organism, in distinction to personal space or extrapersonal space (Rizzolatti, Fadiga, Fogassi, & Gallese, 1997). More specifically, we understand this extension to include a manipulation of physical objects in the environment; a telescope, while vastly altering many features of an organism's capacities and understanding, would not count as extending peripersonal space, in contrast to a tool like a rake or pole. Computer programs such as IDA, the precursor to LIDA, may not have any peripersonal space. Although "embodied" in other ways, since IDA can only sense via network connections and has a physical body made of the computer hardware, it may lack the sense of peripersonal space defined here. (Also see Jackson (2014) for a critical overview of the empirical literature concerning the extension of peripersonal space.)

self, the nature of egocentric and allocentric representations takes center stage. For instance, contrast an agent walking down the street with that same agent sitting on a bench and looking at someone else walk down the street. In these cases, two different event representations will be created. The action link will be the same (walking) while the agent links would be different (self-node vs. other-node).

In contrast to amodal or classical forms of representations, LIDA representations in PAM are grounded in features of the system's body and environment, both internal and external. Recent work on LIDA's spatial representation has been in the direction of "clustering" in spatial memory (Madl, Franklin, Chen, Trappl, & Montaldi, 2016). According to this hypothesis, spatial memory "contains spatial representations of objects in navigation space which are allocentric (world-centered instead of stored in relation to the organism), goal-independent, and are stored long-term" (Madl, et al., 2016, p. 46). The self-node is attached to a place node in response to relevantly strong activation strength (Madl, Franklin, Chen, & Trappl, 2013). The egocentric frame of reference and ability to act on objects in peripersonal space is achieved in the combination of LIDA's Workspace, spatial memory, and the self-node.

To further explain egocentric representations, consider the case of Juan walking to his local grocery store down the street. The geography of the street is represented in a cognitive map. At the level of the LIDA conceptual model, which is more abstract than the cells of Juan's hippocampus, the cognitive map would include place nodes that are represented in LIDA's workspace. The self-node is ultimately attached to whichever place node has the strongest activation, since this is the place where the individual is located at that given time. This attachment of self-node to place node allows Juan to have an egocentric frame of reference in a map that was originally encoded with allocentric representations.

3.2 Minimal Experiential Aspects

Minimal experiential aspects are related to one's sense of ownership (e.g. this action is happening to me) and sense of agency (e.g. I am the one doing this action). Likewise, insofar as a system can be conscious, there will be prereflective experience of a first-person perspective, which includes the impact of various sensory modalities shaping the self/non-self distinction.

Prereflective experience must be conceptually distinguished from preconscious content.[5] Preconscious content concerns the coalitions and structures in the Workspace before a Global

[5] Since LIDA is heavily tied with GWT, a few more words on the relation between different kinds of consciousness are in order. GWT offers an account of functional consciousness (Baars, 1988). For LIDA, the primary purpose of functional consciousness is to filter the Current Situational Model, which includes current sensory information, the memories of various events, and other built structures, for saliency. An equally prominent function of functional consciousness may also be the global broadcast itself. If a conscious broadcast occurs, we take the content of it to be thoughts. A working hypothesis of GWT is that the conscious broadcast will be associated with phenomenal experience. Insofar as LIDA fleshes out GWT computationally and conceptually, one may assume that it should carry a similar assumption. However, while there is an instance where phenomenal consciousness is hypothesized for part of the conscious broadcast in GWT (Franklin & Ramamurthy, 2006, Sect 3.5), the majority of the time the

Broadcast has occurred (Franklin & Baars, 2010). Prereflective experience concerns the lack of a sense of self in conscious contents while engaging in an activity. This sort of experience results from not focusing one's attention on the fact that they are the one doing the action or being acted upon.

Let us first consider an example to help elucidate the differences between sense of ownership and sense of agency. These two commonly run hand in hand. We often feel that we own the acts of which we are an author. I am currently the one typing words onto the page. I am the agent doing the typing and the typing is something that I am doing. We may nevertheless distinguish these two senses conceptually and sometimes in practice. Consider the case of lunging forward vs. being pushed forward. Both cases are things that happen to *me*. When I lunge forward, it is something that I do to myself. In terms of the actions taken, a lunge and a push may result in very similar embodied consequences. The movements of the limbs and body as a whole may even be identical to an outside observer. Yet I am nevertheless only the author of the action in this first case. I have a normal sense of both ownership and agency when I lunge forward. When I am pushed, however, a normal sense of ownership remains, since *I* am who was pushed and who has been moved, with no sense of agency since I was push*ed*

Most of our day-to-day activities take place without reflection. Consider the case of turning a door handle to walk into a new room and contrast that experience with the case of reflecting where you focus your attention on the fact that it is you opening the door handle, as well as being mindful of how the handle feels when being turned. These latter cases are instances of reflective and mindful experience. They concern the way one conceives of themselves in relation to the action. The first case of simply opening the door is an instance of prereflective experience. There is still someone experiencing something it is like to turn the handle, no doubt, but it is not being focused on during the action. Likewise, the fact that I am the one turning the handle is also not focused on in prereflective experience. In reflective experience, I consciously think about me being the one who is acting in whatever ways I am at the time.

In terms of preconscious content, the minimal experiential aspect for LIDA can be implemented by the self-node in PAM. More specifically, the self-node can be recruited from PAM as part of a structure in the Workspace that may not become conscious. As noted in Section 3.1, PAM already has a self-node that can be slotted in a variety of links. A self-node present in a coalition allows for the AA to have a preconscious sense of self/non-self, even in cases where the structure is never part of a coalition that wins the competition for a conscious broadcast (Franklin & Baars, 2010).

---

authors have claimed LIDA has functional consciousness without asserting it has phenomenal consciousness (Franklin, 2003). This latter claim is consistent throughout the development of LIDA and, indeed, it is an open possibility that some LIDA agents may not qualify as phenomenally consciousness. Franklin has suggested that everything besides the phenomenal what-it-is-likeness can be easily captured by functional consciousness. This fact leaves open an important question: to what extent should LIDA claim phenomenal consciousness? How work on the self exactly relates to phenomenal consciousness will be the focus of continuing research.

Prereflective experience is implemented by a LIDA agent when information from the Current Situational Model is broadcast without a self-node. The system as a whole is therefore able to filter, act on, and update itself according to the situation without necessarily having a 'self' involved in the process. Reflective awareness, in contrast, would be implemented when a self-node is part of the coalition whose content gets broadcast.

3.3 Affective Aspects

These aspects are concerned with the manifestation of certain feelings. Features of affective aspects can vary from basic emotions to moods, attitudes, and temperament. Basic emotions, including anger, happiness, sadness, or fear, concern events. Moods last for a longer amount of time, and are generally more sustained experiences. Temperament includes a mix of these features, and usually holds on even longer time scales. Attitudes are the ways in which one engages with the world and various situations within it (Franklin & Ramamurthy, 2006).

LIDA may be uniquely positioned within Artificial General Intelligence research to account for affective aspects of self insofar as it places feelings, including emotions, at its core (Boden, 2016). According to Franklin & Ramamurthy (2006), motivation, feelings, and values are integrally connected for LIDA. Accounting for the implementation of affective aspects in LIDA involves both this general theory of affect, and the specific role for an agent's self-concept (also see Sect. 4.3.2 for more information on the idea of a self-concept).

Considering the ubiquitous role of affect, a complete story of its implementation goes beyond the scope of this paper (see Franklin & Ramamurthy, 2006 for the larger story). Some of its concrete role will be in shaping all aspects of the cognitive cycle. In addition, feeling nodes will be integrally related to coalitions being built from structures in the Workspace. The activation of the feeling nodes with certain valences will impact subsequent actions of the AA (McCall, 2014). In addition, McCall et al. are currently exploring artificial motivation for cognitive software agents within the LIDA model.

One specific way of understanding these aspects concerns Newen and colleagues' pattern theory of emotion (2015). According to this theory, emotions are a pattern of jointly sufficient features. Some of these features will be contingent on the body of the system in question, as well as the general response that these embodied aspects entail. None of these specific embodied features will be implemented beforehand in the LIDA conceptual model. This point notwithstanding, the nature of how feelings have an impact on an organism's mind has been implemented in the LIDA model already (see Franklin & Ramamurthy, 2006).

3.4 Intersubjective Aspects

Some AA develop in social settings with large amounts of intersubjective interaction. Humans often rely on dyadic or triadic engagements early in life - such as interacting with a caregiver (dyadic) or reading the reactions of a caregiver in order to better respond to novel stimuli in the environment (triadic) - that help develop a more robust sense of self. The mirror test (Gallup,

1982) and the introduction of language for some AA (such as *Siri* or Alexa) are also directly related to intersubjective aspects of self.

For example, developmental research has found that infants often appeal to caregivers in order to figure out how to respond to ambiguous stimuli, which could include new objects or other agents in the environment (Gunnar & Stone, 1984). Such behavior is not only limited to infancy, however. The role of peer groups and others in shaping features of an individual's personality, including political choices and aesthetic taste, are important intersubjective aspects of self. In a social nonhuman animal case, infant vervet monkeys learn the proper responses to different predator alarm calls from their mothers (Seyfarth, Cheney, & Marler, 1980). This learning process has been simulated using a LIDA-based agent. (Ait Khayi & Franklin, 2018)

The implementation of these aspects will draw on Semantic Memory, a subset of Declarative Memory, and the perception of other actors and subjects of an event. At the same time, intersubjective aspects will directly relate with the development of attention codelets and structure building codelets. Since these codelets can be said to indirectly shape both Action Selection and schemes in Procedural Memory, the implementation of intersubjective aspects will occur there as well.

3.5 Psychological/Cognitive Aspects

These aspects cover a rather wide range of possibilities, including more robust self-consciousness and conceptually mediated knowledge of the self. Neuroscientists consider these aspects in terms of neural processes, philosophers consider them as part of internal and private sorts of experiences, and all parties in favor of them agree that they underlie our ability to "represent oneself to oneself as oneself."

The implementation of these aspects is an extension of reflective experience. The main difference is their specific content. For reflective experience in general, a self-node will be involved with the winning coalition. The action or experience is therefore something that the agent is aware they are doing. Even passive actions may include elements of agential control and more active processing, such as suggested by accounts of perception that foreground the connection between action and perception (Gangopadhyay & Kiverstein, 2009).

An example of how an apparently passive perceptual process could involve action can be seen in attempting to take part in a conversation in a noisy environment. A person could turn their head in an attempt to hear the speaker better. This would be a clear case of an action. Alternatively, while not necessarily moving any part of their body, the listener could choose to focus their attention by attempting to filter out excess noise and focusing on the voice in question. The choice of mental filtering, however, doesn't just happen to the agent. It is something that the agent may decide to do. Though such a decision may be deliberative (taken consciously), in most cases it will be taken unconsciously, but will be consciously mediated by the incoming salient voice. Just as it is possible to choose the target of perception or tune out some external stimuli by

choice, we suggest that certain sorts of attention shifting processes deserve to be considered actions in the same way as shifts of the head and other limbs are called actions.

These two examples are not the same as psychological or cognitive aspects, however. In contrast to reflective experience in general, psychological/cognitive aspects need to go beyond allowing a person to reference the fact that they are the one doing something. The content of the reference must specifically be about oneself. As a result, in implementation terms, the self-node is either implemented in the agent-link or must be part of multiple links in the same coalition. This requirement is to emphasize the difference of being aware that I am doing something vs. referring to myself as the agent who is doing something. In other words, if we consider focusing on what it is like to feel a door handle as an example of reflective experience, then psychological/cognitive aspects concerns the possibility for one to explicitly think I am the one who feels the door handle.

It is important to note that the ability to "represent oneself to oneself as oneself" does not necessarily require an ability to conceptualize oneself in explicitly linguistic terms. However, it is more robust than the self/non-self distinction found in minimal experiential aspects. For instance, a bacterium may be able to move up a food gradient without having any thoughts or representations that it is the organism in the environment that is moving around the gradient. In contrast, animals that pass the mirror test must do more than the bacterium in the food gradient. Hiding behavior in certain birds may be another example of more robust behavior in this regard (Bugnyar & Kotrschal, 2002).

3.6 Narrative Aspects

Narrative aspects are not just concerned with narratives in the sense of written or linguistic narratives. Some defenders of narrative accounts suggest that selves are often fictions or abstractions (Dennett, 1991), while others suggest that narratives are constitutive of self (Gallagher & Hutto, 2008). There has also been empirical work conducted on the role of narrative and self (Gallagher, 2000). In these cases, narratives are important insofar as they allow a system to make sense of experiences in a certain way.

Some implementations of narrative will utilize language. However, not all narratives are necessarily linguistic in nature. As a result, a narrative may be understood as a sequence of events represented with LIDA's node-link structure. In order to be a successful narrative, these events need to have some sort of relation, yet they don't necessarily need to be coherent in the sense of following a logically rigorous chain of premises towards a conclusion (Launer, 1999). For example, consider looking across the street at a bus station, where a person who is sitting on the bench gets up, walks over to the bus, and steps inside before it takes off. To understand that whole sequence of events under the scope of one overarching event – in this case, the person is catching the bus – requires linking together individual pieces together within a successful narrative about the action.

In order for a narrative aspect of self to be implemented, the self-node must occur within those representations. Part of the implementation of these aspects in LIDA will take place in language use. Other parts of the implementation will rely on PAM, Declarative Memory, and Transient Episodic Memory. Since an important part of narrative aspects is making sense of our actions, they may be an integral part of the story for consolidating Transient Episodic Memory into Declarative Memory either between cognitive cycles or, since cognitive cycles often overlap, during sleep when Transient Episodic Memory is not in the process of receiving new input (Franklin, Baars, Ramamurthy, & Ventura, 2005). Further research into the function of memory consolidation will have an impact on the story of how narrative relates to the consolidation process as well as how different strategies and mechanisms for memory consolidation have an impact on the structure of narratives.

3.7 Extended Aspects

According to Gallagher, following William James, "We identify *ourselves* with stuff we own, and perhaps with the technologies we use, the institutions we work in, or the nation states that we inhabit" (Gallagher, 2013, p. 4 emphasis added). My cell phone is thus not something that I simply use but, rather, something that has become a part of me, at least as much as a toe or foot is part of me if not more.

The various stuff one owns will not be directly implemented in the LIDA model. The specific feelings that one has when they lose those important things, or they are not working properly, on the other hand, can be implemented by LIDA. We suggest that the implementation of these aspects will take place largely in the Current Situational Model, and will involve the general motivational structure of LIDA. This implementation will be integrally related to similar processes for the implementation of affective aspects.

3.8 Situated Aspects

These aspects shape the kinds of selves that we have. The role of familial upbringing and cultural norms, for instance, are of particular note here. Addressing the implementation of these aspects is not about how LIDA can implement them in isolation from the environment and other norms; rather, we understand situated aspects as fundamentally impacting the creation of structures, codelets and action plans in LIDA. In other words, we suggest that one way of understanding situated aspects can be as an individual learning something that gets associated with the self-node.

Learning can introduce new attention and structure building codelets into a system. It also can introduce major shifts in other long term memories, such as PAM, Spatial Memory, Transient Episodic Memory, and Declarative Memory. Spending years reading in a field, perfecting an artistic craft, or working on a new hobby opens up certain possibilities for thinking and acting. What is unique about these situated aspects of self in contrast to learning in general is that they focus on the self. Learning about fashion is an important thing one does over the course of their life, for instance. The idea that I shouldn't wear white after Labor Day is a situated aspect of self.

The idea that green is a fashionable color this year is not. For LIDA, the former will include the self-node while the latter will not.

4. A Conceptual Implementation of a Pattern Theory of Self in LIDA

We now turn our attention to how LIDA can implement the various aspects of the pattern theory of self for its self-model. This will provide a conceptual framework for anyone wanting to implement self in LIDA based agents and similar cognitive architectures. Following Gallagher, we take self to include eight core aspects: minimal embodied aspects, minimal experiential aspects, affective aspects, intersubjective aspects, psychological/cognitive aspects, narrative aspects, extended aspects, and situated aspects.

It is important to reiterate that we are here talking about implementing self in LIDA, not "a" self or "the" self. So the following is not a list of necessary components that must be implemented in a LIDA agent in order for it to have self or a self. What we are doing here is illustrating systems related to various aspects of self that are in place in LIDA that will allow LIDA agents to have a self, depending on the specific implementations of these various systems in a particular LIDA agent. Exploring these eight aspects insofar as they are already present, or can be included, in the LIDA model will have important ramifications for building future LIDA agents.

4.1 Proto-self

The proto-self is generally implemented by a combination of minimal embodied aspects and minimal experiential aspects. However, it is possible that the proto-self can be implemented with as little as minimal embodied aspects. Being embodied, here, does not necessarily preclude purely computational agents from having a proto-self. A computational agent will be realized on hardware of some kind and operate in an environment of some kind, after all (cf. Franklin, 1997). While there is disagreement over whether or not this level of embodiment is enough for certain types of cognitive capabilities and phenomenal consciousness (Boden, 2016, Chapter 6), we take the possibility as an open question that should not be dismissed without further argumentation. Being implemented by one aspect is unique among the different types of selves.

One may worry that including proto-self with just minimal embodied aspects as a kind of self is problematic insofar as it is too weak to qualify as a notion of self. The sense of weak here is meant in the sense of not being part of the cluster concept of self in the first place. According to this worry, our definition of self needs to be something more than the fact that there is a boundary between the system and its environment. This tension is eased when we note that the proto-self was not meant as a stand-alone self. Rather, it is a core set of capabilities that can be built upon to actually implement a self in a system. As a result, the working hypothesis is that all subsequent types of selves in the LIDA self-model will have at least some configuration of the minimal aspects of self in order to be implemented.

4.2. Minimal (core) self: Self-as-subject, Self-as-experiencer, and Self-as-agent

These three kinds of self are implemented with at least minimal embodied aspects and minimal experiential aspects. The specific kind of implementation will focus on the relation between the self-node to the event structure of LIDA. There are also at least two different time scales of self in an immediate instance and self across a developmental process. Depending on the time scale we focus on, different patterns of aspects may be more important.

While the self-as-subject can be realized either reflectively or prereflectively, it is implemented in either case when the self-node is attached to an event representation via the subject link. For self-as-experiencer, the self-node is attached via an agent link. For instance, we can consider the situation where "John heard the bell". In this case, John is slotted into the agent link, heard into the action link, and the bell into the subject link.[6] When the individual in question is addressing their own experience of having heard the bell, we would replace "John" with "I" and thus have a case of a self-as-experiencer. It is important to note that the self-as-experiencer cannot take place from a prereflective case; instead, it will always be either preconscious or reflective. Finally, for the self-as-agent, there will be an event representation where the self-node is attached via the agent link.

4.3 Extended Self

All eight aspects will have different roles to play for different types of extended selves, which include the autobiographical, self-concept, volitional, and narrative selves (See Section 2 above). It is important to note here that the extended self is not the same thing as extended aspects of self. The former captures a specific set of types of selves. The latter includes the external components of an environment that an agent can incorporate in different types of selves.

While the minimal self can be easily accommodated in the current LIDA model, the extended self may require the addition of new processes within existent modules. Some of these processes could include language and story-telling abilities, new structure building codelets, and schemes in Procedural Memory, as well as the implementation of semantics in Declarative Memory. Introducing these different processes will also allow for LIDA to capture additional features of the full eight aspects of self.

4.3.1 Autobiographical Self

The autobiographical self could include all eight aspects of self. However, the most important for it will be narrative aspects and extended aspects. The narrative component is concerned with Transient Episodic Memory and Declarative Memory in the case of events that are composed in a way that contain the self-node. The extended aspects include the objects, tools, and other things in the environment to help sediment the content of one's autobiographical self. Since an AA is

[6] See Section 3.5 for more details on a broader notion of what counts as an action.

situated in and part of their environment, it is quite possible that the extended aspects will have an important role to play in shaping the episodic memory which is central to autobiography.

4.3.2 Self-Concept

In addition to a general role for emotions and affect, another important part of implementing affective aspects involves one's self-concept or what Blackmore calls the selfplex (2000). According to Baars (1988), the self-concept includes one's personal beliefs about themselves. The content of such beliefs do not necessarily need to be explicitly held in mind or reflected upon at all times, although it is most likely the case that most of its content has been reflected upon or held in direct conscious awareness at some time. Within the LIDA model, one likely candidate for the location of the self-concept could be items in Semantic Memory that contain the self-node.

While aspects of one's self-concept may be stable across time, there is also a high amount of context flexibility. An average individual takes up various roles in relation to the different situations. Such role will have a direct relationship with various affective processes and states. The self-concept is implemented in Semantic Memory. This implementation, in turn, will impact the weighting process for action selection, attention, and coalition building.

A fully implemented self-concept will also draw on both a developmental account and a more developed Theory of Mind mechanism (Friedlander & Franklin, 2008). A self-concept, in contrast to the minimal sense of self found in the minimal embodied aspects and minimal experiential aspects, is noticeably more robust. The fact that I take up certain roles rests on understanding myself as distinct from others. In some ways, my ability to understand myself as a certain type or person comes from understanding others as certain types of people.

4.3.3 Volitional self

Some of the most important elements of the volitional self are its affective, extended, and situated aspects. Affect is an essential part of driving an agent to act as well as helping to explain why an agent decided to act in a particular way. An agent iteratively asks the question of 'what should be done next?' While that question should not presuppose a self that is present to be asked, we suggest that any volitional self would be integrally tied into helping answer that question. In addition, the extended and situated aspects of self will help further ground the particular motivations present for an agent's volitional self.

4.3.4 Narrative self

The narrative self may heavily incorporate narrative aspects of self. However, in a similar fashion to the extended self and the extended aspects of self, the narrative self is not exhausted by narrative aspects. It will include features from all eight aspects of self. It is also important to distinguish the narrative self from the autobiographical self.

The narrative self is a broader type of self than the autobiographical self. While both involve a particular ability to string events together in a coherent manner, we don't think there is necessarily a direct entailment between the two. Thus, it is in principle possible to have a narrative self without an autobiographical self and vice-versa. The autobiographical self can be realized in episodic memory with the self-node included in it somewhere. In order to have a narrative self, an agent must be able to tell a story about things. Telling a story often involves the use of language, which necessarily spans across multiple cognitive cycles. Likewise, even non-linguistic narratives will necessarily span across multiple cycles since they draw together different synchronic moments of experience.

5. Conclusion

In this paper we've explored the relationship between the LIDA self-model and the pattern theory of self, as well as suggested the latter serves as a useful theoretical framework for further developing research on modeling self in cognitive architectures. Combining these two has not resulted in any substantive extensions for the original architecture from earlier accounts of the LIDA self-model (Ramamurthy, Franklin, & Agrawal 2012; Franklin et al 2016). However, the use of the pattern theory has helped clarify important theoretical features from the model. It has likewise allowed us to address questions that had remained for the account, such as the relation between the various different types of subselves. The distinct levels of analysis (token, type, and meta-theoretical) from the pattern theory have also helped us bridge the related yet distinct levels of the self-model as a whole, the general subselves like the minimal (core) self or the extended self, and specific subselves like the self-as-experiencer or the volitional self.

Adopting a pattern theory of self has the additional benefit of helping set benchmarks for testing the presence of self in different AA. The importance of these tests is particularly crucial for future development of LIDA agents that are meant to help address different questions about self and disorders of self. Some selves, such as the proto-self, will be built by hand. Others will require different sorts of tests to be developed.

There is no need to run tests for selves that are built by hand. Other selves will be present in light of the ability for an AA to achieve certain ends or take part in certain actions. An AA that can answer the question "why did you do this?" will have a narrative self, for instance, and any AA that has the correct motivational structure will have a volitional self. Still other sorts of selves will have tests that can address sufficient but not necessary conditions. This situation can occur for the autobiographical self or the self-concept. If an AA can answer questions about what happened to it, there is good evidence that the agent has an autobiographical self. It is nevertheless also possible that an AA can have an autobiographical self without the ability to answer questions, so failing a test of that nature cannot preclude the existence of an autobiographical self.

The most difficult selves to test are the three types of minimal (core) self. Part of that difficulty, however, may extend from an assumption that all relevant evidence must be gathered completely from blackbox tests. In the case of built LIDA agents, the creators and testers would know about

relevant internal structures used to implement the self in question, such as structure building codelets that involve a self-node in the relevant event links. Testing these structure building codelets, along with other relevant features of the AA, will allow us to know whether or not that agent has a minimal self of any kind.

Acknowledgements: We would like to thank [Removed for Peer Review]

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of Interest: None

Works Cited

Ait Khayi, N., & Franklin, S. (2018). Initiating language in LIDA: learning the meaning of vervet alarm calls. *Biologically Inspired Cognitive Architectures, 23*, 7-18. doi: 10.1016/j.bica.2018.01.003

Baars, Bernard J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, Bernard J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Science, 6*, 47–52. doi: 10.1016/s1364-6613(00)01819-2

Beni, M. D. (2016). Structural realist account of the self. *Synthese, 193*(12), 3727-3740. doi: 10.1007/s11229-016-1098-9

Blackmore, S. (2000). *The meme machine*. Oxford: Oxford University Press.

Boden, M. A. (2016). *AI: Its Nature and Future*. Oxford, UK: Oxford University Press.

Bugnyar, T., & Kotrschal, K. (2002). Observational learning and the raiding of food caches in ravens, Corvus corax: is it 'tactical'deception? *Animal behaviour, 64*(2), 185-195. doi: 10.1006/anbe.2002.3056

Cochrane, T. (2008). Expression and extended cognition. *The Journal of Aesthetics and Art Criticism, 66*(4), 329-340. doi: 10.1111/j.1540-6245.2008.00314.x

Colombetti, G., & Roberts, T. (2015). Extending the extended mind: the case for extended affectivity. *Philosophical Studies, 172*(5), 1243-1263. doi: 10.1007/s11098-014-0347-3

Damasio, Antonio R. (1999). *The Feeling of What Happens*. New York: Harcourt Brace.

De Haan, S., Rietveld, E., Stokhof, M., & Denys, D. (2017). Becoming more oneself? Changes in personality following DBS treatment for psychiatric disorders: Experiences of OCD

patients and general considerations. *PloS one, 12*(4), e0175748. doi: 10.1371/journal.pone.0175748

Dennett, Daniel C. (1991). *Consciousness explained*. Boston: Little, Brown & Co.

Dings, R., & de Bruin, L. (2016). Situating the self: understanding the effects of deep brain stimulation. *Phenomenology and the Cognitive Sciences, 15*(2), 151-165. doi: 10.1007/s11097-015-9421-3

Franklin, S. (1995). *Artificial Minds*. Cambridge, Ma: MIT Press.

Franklin, S. (1997). Autonomous Agents as Embodied AI. *Cybernetics and Systems, 28*(6), 499–520. doi: 10.1080/019697297126029

Franklin, S. (2003). IDA: A Conscious Artifact? *Journal of Consciousness Studies, 10*, 47–66.

Franklin, S., & Baars, B. (2010). Two Varieties of Unconscious Processes. In E. Perry, D. Collerton, H. Ashton & F. LeBeau (Eds.), *New Horizons in the Neuuroscience of Consciousness* (pp. 91–102). Amsterdam: John Benjamin.

Franklin, S., Baars, B. J., Ramamurthy, U., & Ventura, M. (2005). The Role of Consciousness in Memory. *Brains, Minds and Media, 1*, 1–38.

Franklin, S., & Graesser, A. C. (1997). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents *Intelligent Agents III* (pp. 21–35). Berlin: Springer Verlag.

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., . . . Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures*, 105-130. doi: 10.1016/j.bica.2016.04.003

Franklin, S., & Ramamurthy, U. (2006). Motivations, Values and Emotions: Three sides of the same coin *Proceedings of the Sixth International Workshop on Epigenetic Robotics* (Vol. 128, pp. 41–48). Paris, France: Lund University Cognitive Studies.

Friedlander, D., & Franklin, S. (2008). LIDA and a Theory of Mind. In P. Wang, B. Goertzel & S. Franklin (Eds.), *Artificial General Intelligence 2008* (pp. 137-148). Amsterdam: IOS Press.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Science, 4*, 14–21. doi: 10.1016/s1364-6613(99)01417-5

Gallagher, S. (2011). *The Oxford handbook of the self*: Oxford University Press.

Gallagher, S. (2013). A pattern theory of self. *Frontiers in Human Neuroscience, 7*. doi: 10.3389/fnhum.2013.00443

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. *The shared mind: Perspectives on intersubjectivity, 12*, 17-38. doi: 10.1075/celcr.12.04gal

Gallup, G. (1982). Self-awareness and the emergence of mind in primates. *American Journal of Primatology, 2*, 237–246. doi: 10.1002/ajp.1350020302

Gangopadhyay, N., & Kiverstein, J. (2009). Enactivism and the unity of perception and action. *Topoi, 28*(1), 63-73. doi: 10.1007/s11245-008-9047-y

Gunnar, M. R., & Stone, C. (1984). The effects of positive maternal affect on infant responses to pleasant, ambiguous, and fear-provoking toys. *Child Development*, 1231-1236. doi: 10.2307/1129992

Jackson, G. B. (2014). Skillful action in peripersonal space. *Phenomenology and the Cognitive Sciences, 13*(2), 313-334. doi: 10.1007/s11097-013-9301-7

Kyselo, M. (2014). The body social: an enactive approach to the self. *Frontiers in Psychology, 5*, 986. doi: 10.3389/fpsyg.2014.00986

Launer, J. (1999). Narrative based medicine: A narrative approach to mental health in general practice. *Bmj, 318*(7176), 117-119. doi: 10.1136/bmj.318.7176.117

Madl, T., Baars, B. J., & Franklin, S. (2011). The Timing of the Cognitive Cycle. *PLoS ONE, 6*(4), e14803. doi: 10.1371/journal.pone.0014803

Madl, T., Franklin, S., Chen, K., & Trappl, R. (2013). Spatial Working Memory in the LIDA Cognitive Architecture. In R. West & T. Stewart (Eds.), *Proceedings of the 12th International Conference on Cognitive Modelling* (pp. 384-390). Ottawa, Canada: Carleton University.

Madl, T., Franklin, S., Chen, K., Trappl, R., & Montaldi, D. (2016). Exploring the structure of spatial representations. *PLoS ONE*. doi: 10.1371/journal.pone.0157343

McCall, R. (2014). *Fundamental Motivation and Perception for a Systems-Level Cognitive Architecture.* PhD Thesis, University of Memphis, Memphis, TN USA.

McCall, R., Franklin, S., & Friedlander, D. (2010). *Grounded Event-Based and Modal Representations for Objects, Relations, Beliefs, Etc.* Paper presented at the FLAIRS-23, Daytona Beach, FL.

Newen, A., Welpinghus, A., & Juckel, G. (2015). Emotion recognition as pattern recognition: the relevance of perception. *Mind & Language, 30*(2), 187-208. doi: 10.1111/mila.12077

Ramamurthy, U., & Franklin, S. (2011). *Self System in a model of Cognition.* Paper presented at the Machine Consciousness Symposium at the Artificial Intelligence and Simulation of Behavior Convention (AISB'11, University of York, UK.

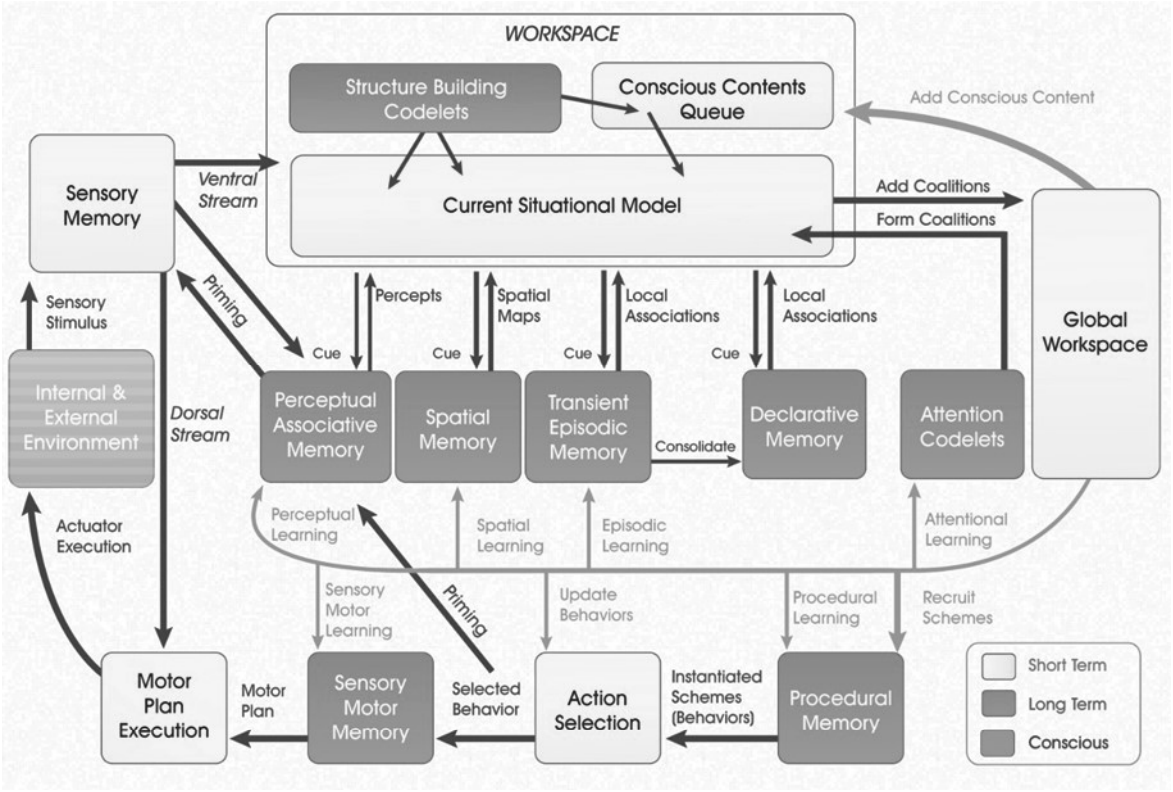
Ramamurthy, U., Franklin, S., & Agrawal, P. (2012). Self-system in a model of cognition. *International Journal of Machine Consciousness, 4*(02), 325-333. doi: 10.1142/s1793843012400185

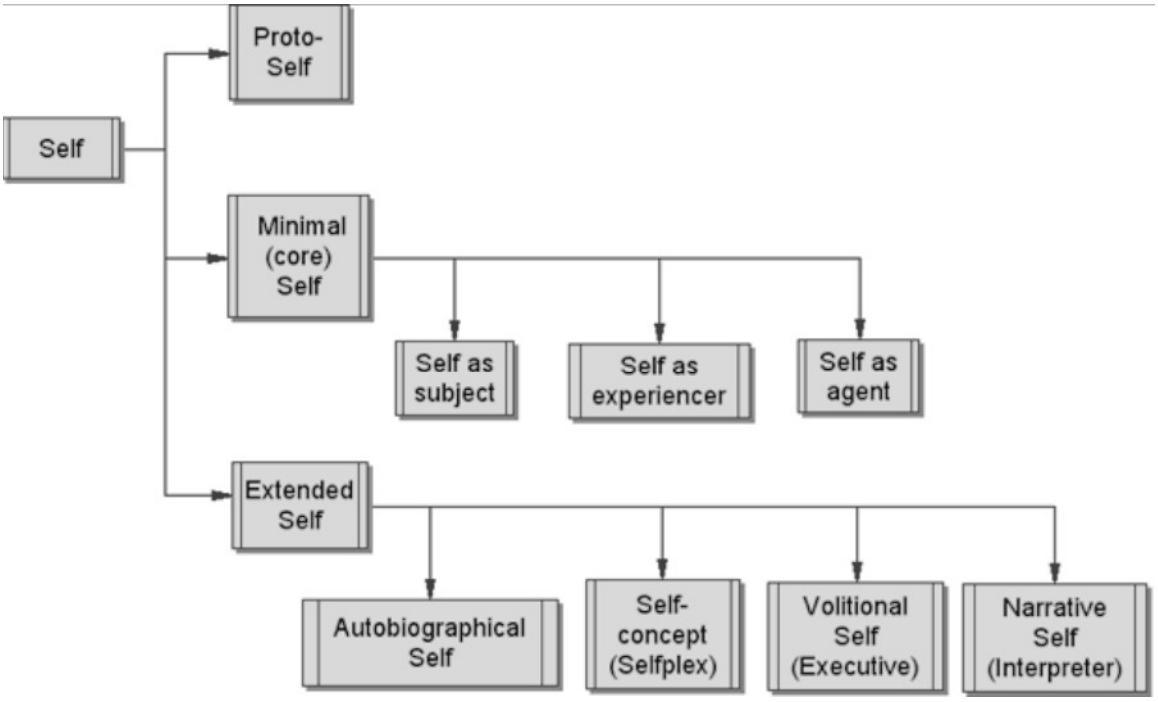
Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (1997). The space around us. *Science, 277*(5323), 190-191.

Seyfarth, R., Cheney, D., & Marler, P. (1980). Monkey responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication. *Science, 210*(4471), 801-803.

Sloman, A. (2002). Architecture-based conceptions of mind *In the Scope of Logic, Methodology and Philosophy of Science* (pp. 403-427): Springer.

Wittgenstein, L. (1953). Philosophical Investigations, ed. by Anscombe GEM, Rhees R, von Wright GH, trans. by Anscombe GE M: Oxford, Basil Blackwell.


WORKSPACE
Structure Building Codelets
Conscious Contents Queue
Add Conscious Content
Sensory Memory
Ventral Stream
Current Situational Model
Add Coalitions
Form Coalitions
Global Workspace
Sensory Stimulus
Priming
Percepts
Spatial Maps
Local Associations
Local Associations
Cue
Cue
Cue
Cue
Internal & External Environment
Dorsal Stream
Perceptual Associative Memory
Spatial Memory
Transient Episodic Memory
Consolidate
Declarative Memory
Attention Codelets
Actuator Execution
Perceptual Learning
Spatial Learning
Episodic Learning
Attentional Learning
Sensory Motor Learning
Priming
Update Behaviors
Procedural Learning
Recruit Schemes
Motor Plan Execution
Motor Plan
Sensory Motor Memory
Selected Behavior
Action Selection
Instantiated Schemes (Behaviors)
Procedural Memory
Short Term
Long Term
Conscious


Self
Proto-Self
Minimal (core) Self
Self as subject
Self as experiencer
Self as agent
Extended Self
Autobiographical Self
Self-concept (Selfplex)
Volitional Self (Executive)
Narrative Self (Interpreter)

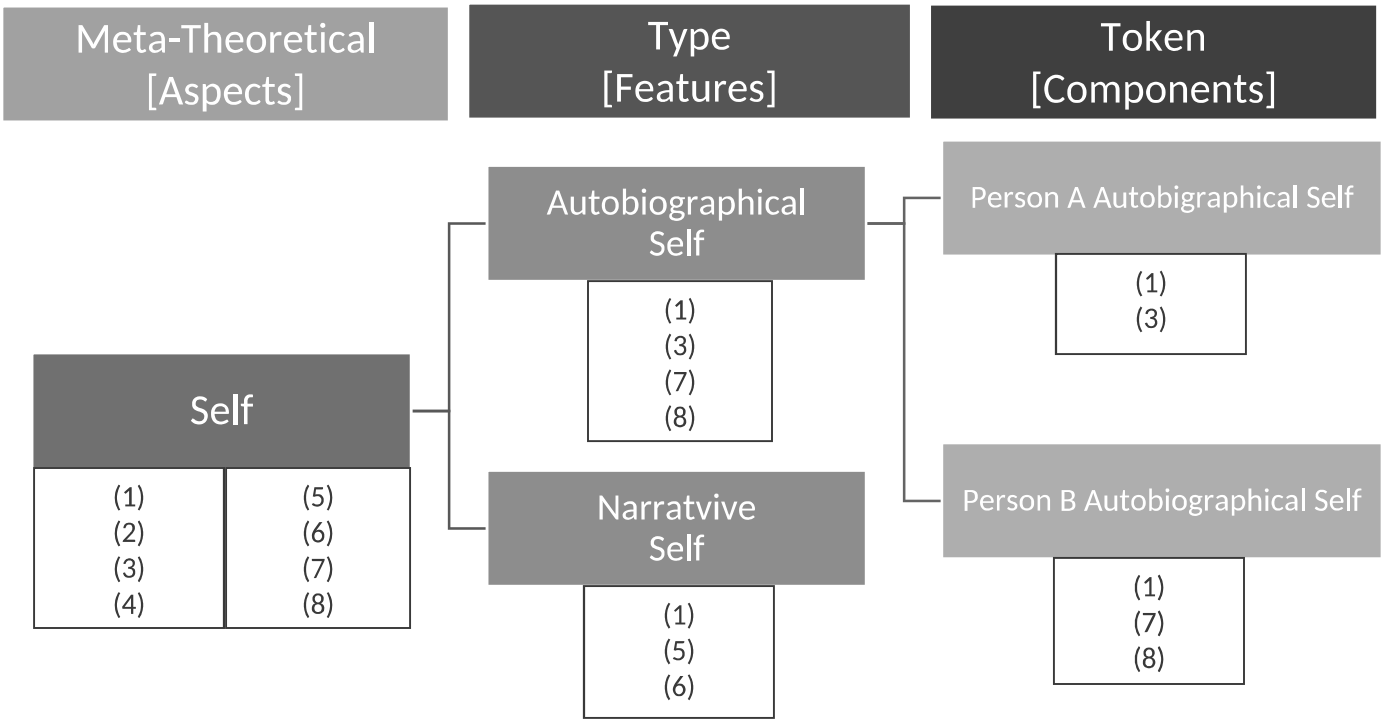


Figure 3. The Three Levels of the Pattern Theory of Self

This figure shows how the three different levels of the pattern theory of self are related. Each *type* of self will have specific features that are drawn from the possible aspects of self as specified at the *meta-theoretical* level. At the same time, since the relevant patterns are jointly sufficient clusters of properties, it is possible that different *token* selves of the same type will have various forms of realization. Important details of the specific dynamics and conceptual implementation of these aspects, features, and components are not captured here but will be explored in more detail below (see section 4 in particular).

April 23, 2019

RE: Conflict of Interest Form for "The Pattern Theory of Self in Artificial General Intelligence: A Theoretical Framework for Modeling Self in Biologically Inspired Cognitive Architectures" by Kevin Ryan, Pulin Agrawal, and Stan Franklin

To Whom It May Concern,

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author.

Yours Sincerely,

Kevin Ryan
Pulin Agrawal
Stan Franklin