Cognitive Systems Research xxx (xxxx) 1–12

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com

Modeling interactions between the embodied and the narrative self: Dynamics of the self-pattern within LIDA

Alexander Hölken [a, *], Sean Kugele [b], Albert Newen [a], Stan Franklin [b, 1]

[a] Institute of Philosophy II, Ruhr-Universität Bochum, Universitätsstrasse 150, D-44801 Bochum, Germany
[b] Computer Science Department and Institute for Intelligent Systems, The University of Memphis, 3720 Alumni Ave, Memphis, TN 38152, United States

ARTICLE INFO

Action editor: Alexei V. Samsonovich

Keywords:
Biologically Inspired Cognitive Architecture
Artificial General Intelligence
Cognitive Modeling
Autonomous Agents
Self-directed Learning
Pattern Theory of Self
Embodied Self
Narrative Self

ABSTRACT

Despite lacking a generally accepted definition, Artificial General Intelligence (AGI) is commonly understood to refer to artificial agents possessing the capacity to build up a context-independent understanding of itself and the world and to generalize this knowledge across a multitude of contexts. In human agents, this capacity is, to a large degree, facilitated by processes of *self-directed learning*, during which agents voluntarily control the conditions under which episodes of learning and problem solving occur. Since self-directed learning depends on the degree of knowledge the agent has about various aspects of themselves (their bodily skills, their learning goal, etc.), an AGI implementation of this type of learning must build on a theory of how this self-knowledge is actualized and modified during the learning process. In this paper, we employ the *pattern theory of self* in order to characterize different aspects of an agent's self that are relevant for self-directed learning. Such aspects include agent-internal cognitive states such as thoughts, emotions, and intentions, but also relational states such as action possibilities in the environment. Combinations of these aspects form a characteristic pattern, which is unique to each individual agent, with no one aspect being necessary or sufficient for the individuation of that agent's self. Here, we focus on the interdependence of narrative and embodied aspects of the self-pattern, since they involve particularly salient challenges consisting in conceptualizing the interaction between propositional and motor representations.

In our paper, we model the reciprocal interaction of these aspects of the self-pattern within an individual cognitive agent. We do so by extending an approach by Ryan, Agrawal, & Franklin (2020), who laid the groundwork for the implementation of the pattern theory of self in the LIDA (Learning Intelligent Decision Agent) model. We describe how embodied and narrative aspects of an agent's self-pattern are realized by patterns of interaction between different LIDA modules over time, and how interactions at multiple temporal scales allow the agent's self-pattern to be both dynamically variable and relatively stable. Finally, we investigate the implications this view has for the creation of artificial agents that can benefit from self-directed learning, both in the context of deliberate planning and adaptive motor execution.

1. Introduction

Artificial General Intelligence (AGI) entails the capacity to "carry out a variety of different tasks in a variety of different contexts, generalizing knowledge from one context to another, and building up a context and task independent pragmatic understanding of itself and the world" (Goertzel & Pennachin, 2007, pg. 74). Furthermore, we contend that generally intelligent systems will likely by autonomous agents[1] that can recognize relations between themselves and their environment, and reason about what to do (and how to do it) based on their self-knowledge. As such, implementing a sense of body and self appears to be a fundamental prerequisite for constructing generally intelligent systems. Knowledge about self-related states (e.g., one's current goals or one's current pose) is a requirement for cognitive capacities such as *meta*-cognition (Cox, 2005), and the routine and adaptive execution of sets of movements (e.g., during skillful interaction within familiar envi-

* Corresponding author.
E-mail addresses: alexander.hoelken@rub.de (A. Hölken), skugele@memphis.edu (S. Kugele), albert.newen@rub.de (A. Newen), franklin@memphis.edu (S. Franklin).

[1] Stan Franklin passed away on January 23rd, 2023 at the age of 91. He was a prolific writer, a superb collaborator, and a patient mentor. His contributions to this paper and to cognitive science were many, and we are deeply saddened that he did not live to see the publication of our joint work. He will be deeply missed.

[1] Franklin and Grasser (1997) define an autonomous agent as "a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (pg. 25).

https://doi.org/10.1016/j.cogsys.2023.03.002
Received 15 December 2022; Received in revised form 28 February 2023; Accepted 5 March 2023
1389-0417/

ronments). These and other cognitive capacities rely on the agent possessing a complex sense of self that spans multiple different domains, each of which plays a key role in determining how well the agent is able to adapt known information or skills to new environments. Such a pluralist account of self has strong implications for building AGI, in particular if these systems are to be capable of "self-directed learning" (Knowles, 1975) – that is, capable of controlling their own learning experience based on knowledge about their own abilities, limitations, and goals (Gureckis & Markant, 2012; Morris, 2019). In the Philosophy of Mind, Shaun Gallagher and Albert Newen have proposed the so-called *pattern theory of self*, which provides a theoretical framework for capturing the multifaceted nature of an individual agent's self (Gallagher, 2013; Newen, 2018; Gallagher & Daly, 2018).

The pattern theory explicitly rejects the traditional conceptualization of "the self" as a unitary entity with certain properties, but defines the self as a pattern of cognitive aspects of agents which govern their thought and behavior. Such aspects include agent-internal cognitive states such as emotions, intentions, and skills, but also relational states between the agent and their environment (Gallagher, 2013). Aspects of both kinds combine to form an agent-specific pattern of states, processes and relations, including experiential, affective, psychological, narrative, normative, and many other factors of cognitive experience. As these factors are closely interrelated, an intervention that affects one factor will involve modulations in the other factors. Adjustments in one aspect, above a certain threshold, will lead via dynamical interactions to changes in others. For example, very basic aspects of self-experience, such as the sense of agency, can be modulated by more complex, relational aspects, such as social normative factors that involve culture, gender, race, health, etc. (Gallagher & Daly, 2018)

With its characterization of the self as a pattern of dynamically interacting aspects, the pattern theory provides useful conceptual tools for implementing different varieties of self-directed learning within artificial autonomous agents. For instance, how an artificial agent possessing general intelligence can control their own learning experience depends on the different types of knowledge they have about themselves: Their immediate goals for the discrete learning situation, the overall narrative goals that the learning situation contributes towards, their own cognitive and physical abilities, and – in cases of social learning – the relation between their learning progress and performance and that of their peers.[2] Understanding these different types of self-knowledge as knowledge about aspects of one's self-pattern allows us to account for a wide variety of situations in which a particular aspect (e.g., a narrative goal) plays a central role in guiding the agent's learning behavior, and how a change to that aspect affects different aspects of their self-pattern over time.

In a recent analysis of the interaction between aspects of self-patterns over time, Newen (2018) has argued that an agent's self-pattern is distributed over both short and long timescales, allowing it to be dynamically modified during concrete situations of embodied interaction while staying relatively stable over longer periods of time. According to this picture, aspects of a self-pattern influence each other across different time scales, with the agent's behavior in a concrete situation based on cognitive dispositions of the narrative self (e.g., narrative deliberation) causing long-term dispositions of the embodied self to develop in different ways (e.g., stabilizing a behavioral pattern).[3] Dings (2018) provides concrete examples of this mutual influence by investigating the temporal relationship between narrative self-programming and the development of sensitivities to affordances in the agent's environment. Building on these attempts to flesh out the notion of *aspects of the self-pattern* and the temporal relation between them, we provide an account of how these aspects can be realized within autonomous agents, and how they interact with each other both over short and long timescales.

In order to achieve this goal, this paper is divided into two main parts: We begin with a coarse-grained, functional description of the temporal dynamics of aspects of an agent's self-pattern and how they develop and interact over time, based on the work of Newen (2018) and Dings (2018, 2021). In the second half of the paper, we will make use of the Learning Intelligent Decision Agent (LIDA) cognitive architecture in order to model these functional considerations within a more detailed framework for modeling cognitive agents capable of autonomous learning (Kugele & Franklin, 2021). We decided to use the LIDA architecture for this model, as its organizational structure is biologically inspired and aligns with some core concepts of the pattern theory. In addition, LIDA aims at being compatible with a unified theory of cognition (Newell, 1994; Kotseruba & Tsotsos, 2018), thus providing adequate tools for modeling self-patterns in any agent with general intelligence, whether artificial or human. In particular, we believe that some of the fundamental concepts employed by the pattern theory (e.g., *aspects of a self-pattern*) can be clarified by identifying them with entities within the LIDA model.[4] Furthermore, we aim to explicate the notion of "dynamical interactions" (Gallagher & Daly, 2018) between aspects of a self by modeling the influence of one aspect on another. For the purpose of brevity, we limit the scope of our investigation to *embodied* and *narrative aspects*, but our intention is to provide an explanation that can serve as the basis for a generalized account of the temporal structure of interaction for any aspect of an agent's self-pattern. We then wrap up by discussing the general implications of this temporal structure of aspect interaction for modeling artificial agents capable of self-directed learning.

2. The embodied and the narrative self

Our functional description of the interaction between embodied and narrative aspects of the self should be understood in the context of Gallagher's pattern theory of self. However, our definitions of the two aspects we are interested in differs slightly from his: Here, the term *embodied self* refers to self-aspects that are based on non-semantic self-representations (Newen & Vogeley, 2003) such as a body schema and a point of perception from which the environment is experienced. It also encompasses learned behaviors and skills that Gallagher and Daly call "behavioral aspects" (Gallagher & Daly, 2018). In general, we will use "embodied self" as an umbrella term which refers to all non-semantic aspects of the self that support agents in meeting the requirements that the environment places on their bodily activity. In contrast, *the narrative self* is constituted by self-related narratives and cognitive components that create and modify them. Narratives can be characterized as sets of actions and events that are temporally ordered and semantically structured, some of which are self-narratives, because they include a self-representation. These are often invoked when an explicit evaluation of one's past or future action is required, such as during conscious deliberation (Gallagher & Daly, 2018). Finally, we understand both the embodied and the narrative self as aspects of the agent as a whole, and not of a stable, unitary entity that is the "seat" of the self – a view that coheres with the absence of a self-module within LIDA.

2.1. Situational influence of the narrative self on the embodied self

The narrative self is essentially constituted by self-defining narratives, i.e., those which are relevant to how agents see themselves and want others to see them. To understand how *self-narratives* can influ-

[2] Not all of this knowledge needs to be consciously available to the agent for it to influence the process of learning.

[3] We use the term *disposition* to refer to an agent-internal (short- or long-term) tendency to produce certain behaviors.

[4] Gallagher (2013) emphasizes the fact that he does not want the word "pattern" to simply refer to patterns of neural activity, but never specifies which elements of an agent (and perhaps their environment) can realize self-patterns.

ence cognitive processes such as deliberation and problem solving, consider an autonomous agent named Alice. On a given Sunday morning, Alice might deliberate whether to take the day off or to go to her study in order to work on a paper that needs to be finished. A contextually relevant self-narrative may become part of this deliberative process, e.g., one's self-narrative as a diligent academic who finishes her papers before their deadlines. These thoughts may be closely connected with long-term aspects of the embodied self, for instance a habit of working on Sunday mornings that Alice has formed during her years of working as an academic. The existence of this habit biases Alice towards certain activities on Sunday mornings, e.g., by making it very likely for her to enter her study first thing in the morning.[5] Now, suppose that Alice has a young son, who she has promised to go to a soccer game with on the following Sunday. If Alice has a narrative goal of being a good mother, she might make the conscious decision to *not* follow her usual behavioral routine next Sunday, and instead gather the soccer gear for her son. In this case, the challenge for Alice consists in modifying the habit to go to her study every Sunday, since this is a steady part of her behavioral routine and in line with another part of her self-narrative (e.g., "I am someone who takes their work seriously"). On the one hand, robustly anchored habits are not easy to change: Thus, just the thought not to go to her study next Sunday usually does not do the trick. She needs to overcome the behavioral routine of going to her study, which is firmly anchored by a multitude of factors, such as a desire to keep up her work routine, her awareness of social pressure from her academic peers, and others. These factors are core components of Alice's self-narrative of being a hard worker. As such, a conflict between them and Alice's intention to change her habits may eventually lead to a moment of narrative deliberation. This occurs when Alice faces a narrative dilemma, as she realizes that the consequences of two of her self-narratives ("I want to be a good mother" and "I am someone who takes their work seriously") conflict. Thus, she needs to both *deliberate* about which self-narrative is more important to her and then, if that decision conflicts with an already established behavior, find a way to "override" it. In terms of the pattern theory, this means that *her narrative self needs to enact a top-down influence on her embodied self* which consists of the long-term adjustment of dispositions to instantiate certain behaviors, given some context. In Alice's case, this adjustment is the result of narrative deliberation. We will provide a model of this process of narrative deliberation and a resultant change in behavior in Section 4.1.

There are three important features we want to highlight using this functional characterization of the top-down influence: *First*, the mere thought of wanting to change one's behavioral routines is usually not sufficient for actually changing them. This is because these routines are self-reinforcing and anchored within the situations that bring them about, for instance, one's tendency to go to their study right after one has finished eating breakfast. *Second*, the top-down influence can be best described as long-term adjustments of dispositions to behave in a certain way. *Third*, we can describe the influence of the narrative self on the embodied self without presupposing a unitary, non-variable self. We only need to consider the temporal patterns of interaction between the relevant self-narratives and established behavioral routines. Although we presuppose that agents experience themselves as a unitary self within a given situation, this unitary experience is rather variable and in principle open for modification, both in its short-term aspects, which may vary strongly during a given moment in time, and in its long-term aspects, which largely remain stable across long periods of time. However, we do not need to presuppose the existence of a self-entity as part of the cognitive system: What allows us to explain the top-down influence of self-defining thoughts on concrete activities are the relevant self-informational aspects of the cognitive system (the agent) and their systematic interconnections. Let us now see how this applies to cases of the embodied self influencing the narrative self.

2.2. Situational influence of the embodied self on the narrative self

Imagine a non-sportive person who starts to play tennis due to a casual invitation of a friend. Accompanying her friend frequently changes her into a fanatic tennis player who starts to train intensely. This, in combination with some successful matches, modifies her self-defining thoughts in such a way that they now involve playing tennis as an important activity and being sportive as a characteristic feature of herself. This process starts with the situational thought that she wants to arrange a joint activity with her friend once a week and discovers that the best fit is to accompany her to the tennis club. Her friendship then motivates her to play tennis every Friday evening and thereby learn a new behavioral routine. The regular practice transforms this into a stable behavioral routine, i.e., a new habit to play tennis on Friday evenings. This habit influences the agent's way of thinking about herself within concrete situations, such as winning tennis matches: For instance, her thought of being a good tennis player may be reinforced after she has won a match. These regular situational experiences can then transform or create new, persistent self-narratives, such as being a successful tennis player. In terms of the pattern theory, this means that *the embodied self can enact a bottom-up influence on the narrative self* by affecting long-term dispositions to form certain self-defining content. This influence is the result of a series of short-term adjustments of embodied habits within concrete situations (Fig. 1).

The important features of the process of bottom-up influence are analogous to those of the top-down influence discussed earlier: *First*, the mere occurrence of a new behavior is usually not sufficient for new self-narratives to emerge or existing ones to change. This is because there is no immediate influence of the embodied self in the form of a new type of situational behavior on the agent's long-term dispositions to form self-narratives. Rather, new self-narratives arise from the repeated instantiation of certain behaviors (behavioral routines), which we can think of as typically *habituated dispositions* that constrain and modulate self-defining narratives. *Second*, the bottom-up influence can be best described as *the long-term adjustments of dispositions* which trigger the formation of new self-defining contents. *Third*, we can describe the influence of the embodied self on the narrative self without presupposing a stable non-varying entity or self. We only need to involve relevant short- and long-term aspects of both kinds of self-representations. In conclusion, the different aspects of the self-pattern and their systematic interactions allow us to explain both the top-down influence of self-narratives on concrete embodied activities and bottom-up influence of the latter on self-narratives by reference to the creation of new behavioral routines and self-defining contents. With this coarse-grained functional characterization in mind, we now turn to LIDA in order to model this reciprocal interaction between the embodied and the narrative self within a more detailed cognitive architecture.

3. The LIDA cognitive model

Learning Intelligent Decision Agent (LIDA; Franklin et al., 2016) is a systems-level, biologically inspired cognitive architecture that models natural minds (human and non-human) and guides the construction of artificial minds (e.g., intelligent software). LIDA is composed of a set of asynchronously interacting modules and processes (see Fig. 2), which collectively give rise to cognition. Cognition, in this sense, includes all mechanisms of mind, including (but not limited to) perception, motivation, attention, action selection, motor control, learning, language, mental simulation, and sense of body and self. Thus, LIDA can be seen as one path to modeling and implementing *generally* intelligent systems,

[5] We use the term "constraint" in the same sense as the constraints-led approach to skill acquisition. In brief, constraints are limits to which behaviors or actions an agent can instantiate, and how they are instantiated. Constraints can be both agent-internal and environmental. For a more detailed introduction, see Button et al. 2021.

	To Embodied Self	To Narrative Self
From Embodied Self	**Online action control**: Short-term adjustments of embodied behaviors and skills based on environmental challenges	**Modification of narratives**: Long-term adjustments of dispositions to form new self-narratives
From Narrative Self	**Self-programming:** Long-term adjustments of dispositions to instantiate certain behaviors	**Narrative deliberation**: Short-term adjustments of distal goals, based on concrete narrative challenges

Fig. 1. Overview of the ways that the embodied and the narrative self can reciprocally influence each other. For instance, a person may be faced with a concrete situation in which their spouse threatens to leave them unless they quit smoking. This may start a process of *narrative deliberation* about what is more important to them: Smoking or their spouse. Coming to the conclusion that their spouse is more important, they then set a distal goal to quit smoking via *self-programming*. This goal may then become part of a new or existing self-narrative if it is strong enough to affect existing behavioral dispositions, e.g., that of buying cigarettes every morning. This, in turn, may lead to the creation of a new narrative of being a non-smoker.

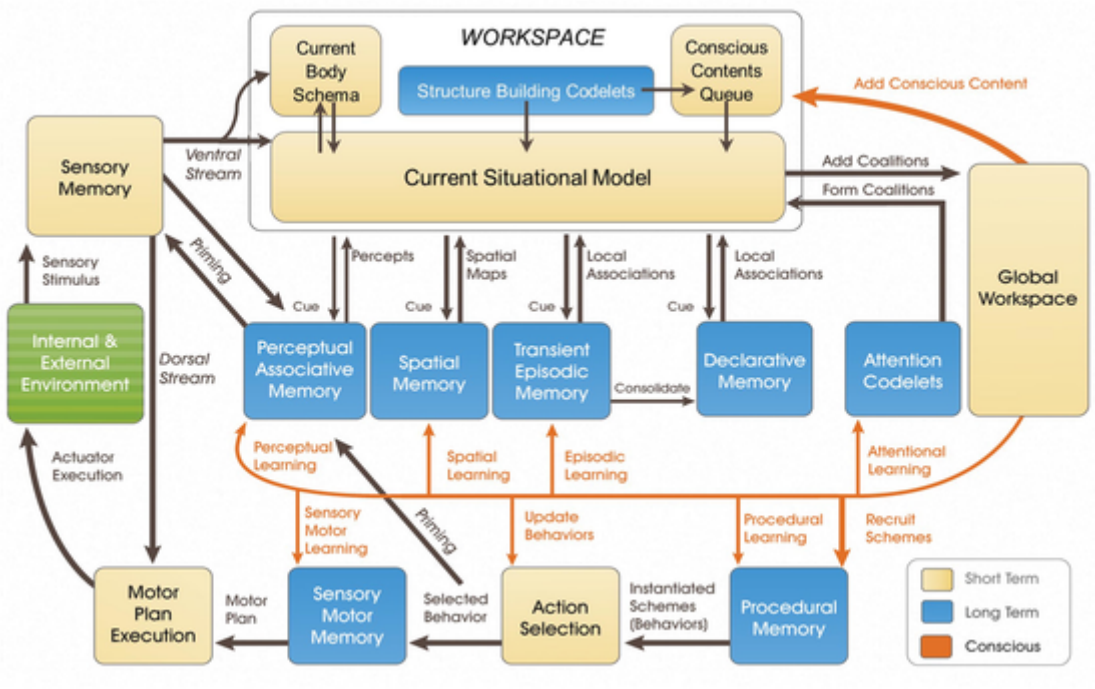

Fig. 2. The LIDA cognitive cycle.

though a great deal of work is needed before that goal can be realized in practice.

3.1. The LIDA cognitive cycle

All cognitive activities occur in LIDA within, or as a result of, a series of potentially overlapping *cognitive cycles*. Cognitive cycles can be thought of as "cognitive atoms" from which higher-order cognitive functions such as deliberation, reasoning, imagination, and problem solving are composed. LIDA's cognitive cycle is analogous to the "action-perception cycle" referred to by many psychologists and neuroscientists (Cutsuridis, Hussain, & Taylor, 2011; Dijkstra, Schöner, & Gielen, 1994; Fuster, 2002, 2004; Neisser, 1976), with each cycle typically lasting between 200 and 500 ms in humans (Madl, Baars, & Franklin, 2011). Conceptually, LIDA's cognitive cycles can be further divided into three phases: perception and understanding, attention, as well as action and learning (Franklin et al., 2016).

During the perception and understanding phase, an internal model – called the Current Situational Model (CSM) – is assembled in LIDA's Workspace that represents an agent's *preconscious*[6] understanding of its current environmental state. This phase begins when environmental stimuli *activate*[7] low-level feature detectors in LIDA's Sensory Memory module. The resulting sensory representations (which are typically im-

[6] LIDA implements a significant portion of the Global Workspace Theory (GWT) of consciousness (Baars, 1988; Baars, Geld, & Kozma, 2021). Consequently, LIDA's representations can be categorized as being either conscious or unconscious at a given moment. Unconscious representations are "preconscious" if an agent could potentially become conscious of them in the future, or "never conscious", if they are permanently excluded from the Global Workspace (see Franklin & Baars, 2010; Baars, Franklin, & Ramsøy, 2013).

[7] Nearly all of LIDA's representations and processes have associated activation-related parameters (Kugele & Franklin, 2020). The precise meaning of these activations is often module-specific, but they can be broadly categorized as falling into one of three types: *Base-level activation* decay relatively slowly and are updated based on an agent's conscious experiences. *Current activations* decay relatively quickly and generally reflect transistory, module-specific notions of "relevance". Finally, *total activations* are typically derived from other activation-related parameters.

plemented as non-symbolic, multi-modal analogs of environmental stimuli) are used to activate representations in Perceptual Associative Memory (PAM). These include higher-level feature detectors and perceptual representations corresponding to objects, entities, and events. PAM representations that receive sufficient activation are integrated into LIDA's CSM as percepts. In addition to identifying the objects, entities, and events that occur in an agent's environment, percepts can also include an agent's affective appraisals of its current situation. These may incorporate an agent's immediate hedonic responses to environmental stimuli (i.e., feelings of "liking" or "disliking") as well as its desires and aversions to real or imagined events (see McCall, Franklin, Faghihi, Snaider, & Kugele, 2020 for a detailed introduction to LIDA's affective and motivational system). Such desires and aversions are typically based on a history of environmental interactions, and their magnitudes are quantified by a parameter associated with each PAM representation called *incentive salience*.

Content in the CSM is continually updated based on incoming environmental stimuli, situation-relevant percepts, *cued* long-term memories (e.g., episodes, narratives, and cognitive maps), and content constructed by *structure building codelets*. Structure building codelets are special-purpose processes that continually monitor LIDA's Workspace for content matching their concerns, and generate or modify content in the CSM as a result. For instance, a structure building codelet monitoring for causal relations might create a "causality link" between event structures for "Lou kicks the ball" and "the dog runs after the ball." In this particular case, a structure building codelet might take these event structures from LIDA's Conscious Contents Queue (CCQ) – a submodule of LIDA's Workspace – which maintains a temporally ordered data structure containing an agent's recent "conscious" broadcasts from previous cognitive cycles (see Snaider, McCall, & Franklin, 2012). Upon retrieving these event structures from the CCQ, this structure building codelet could then create a new structure containing these two events connected by a causality link. This new composite event structure could then be interpreted as, "Lou kicking the ball **caused** the dog to run after the ball.".

During the attention phase, *attention codelets* – special-purpose processes similar to structure building codelets – scan LIDA's CSM looking for preconscious content matching their individual concerns. These concerns (matching criteria) can be very general (e.g., structures with high activation or incentive salience) or highly specific (e.g., structures representing loud noises, novel objects, or surprising events). If such content is found, attention codelets may form *coalitions*[8] that contain that content. The resulting coalitions are then sent to LIDA's Global Workspace where they compete in a winner-take-all competition for inclusion in LIDA's conscious broadcast. This competition is based solely on the coalitions' activations, which are derived from the features of their incorporated structures. These features can include their activations (e.g., situational relevance or the intensity of "feelings"), incentive saliences (e.g., desirability), and how well those structures matched the attention codelets' concerns. Thus, the competition in LIDA's Global Workspace functions as a saliency filter that orients an agent's attention towards the most important, urgent, or relevant (etc.) structures in the CSM. Structural salience is based, in part, on an agent's current needs, activities, and goals. At the end of the attention phase, the winning coalition's content is broadcast to all of LIDA's modules and processes, thereby becoming consciously accessible. This initiates the action and learning phase of the cognitive cycle.

During the action and learning phase, each module and process selects content from the conscious broadcast based on their own needs. This initiates LIDA's many module- and process-specific learning mechanisms (see Kugele & Franklin, 2021), and the selection and eventual execution of actions. Prior to action selection, *schemes* in LIDA's Procedural Memory module are activated based on the content in the global broadcast. A scheme is a data structure containing a situational *context*, a primitive or composite *action*, and an expected *result*.[9] It functions as a unit of procedural knowledge that specifies what might happen if its action were executed in a given context. A scheme also has a *base-level activation* quantifying the likelihood that its result would occur were its action to be taken when its context is satisfied. Schemes receiving sufficient activation are instantiated as *behaviors* and sent to LIDA's Action Selection module. Action Selection chooses one such behavior, and sends it to LIDA's Sensory Motor System (SMS) for execution. The SMS instantiates an appropriate *motor plan* for the behavior, which is then executed through a process of "online control" (Dong & Franklin, 2015). During the action and learning phase of *subsequent* cognitive cycles, Procedural Memory may update the base-level activations of applicable schemes whose actions have been executed. These updates are based on an agent's observations of their environment following action execution. Such schemes are "reinforced" (i.e., their base-level activations are increased) when an agent's observations (i.e., content in the conscious broadcast) match the schemes' predicted results.

3.2. LIDA and the pattern theory of self

LIDA features a multitude of disparate memory modules, each supporting different types of knowledge structures and processes. This feature of LIDA makes it especially relevant to efforts aimed at conceptualizing complex, but nuanced theories of cognition and self, such as Gallagher's pattern theory (Ryan, Agrawal, & Franklin, 2020). For the sake of brevity, we will focus on those memory modules that are relevant to the question of how *narrative* and *behavioral aspects of the self-pattern* can be modeled in LIDA. They are listed in the following table.

	Characterization in LIDA	Relevance for the Pattern Theory
Procedural Memory	Stores *schemes* consisting of a (situational) context, an action, and a result. Schemes also have base-level activations that quantify the likelihood that their results will occur, when their actions are taken in a given context.	Memory of basic actions, embodied skills and behavioral routines (habits). LIDA's schemes can be thought of as part of an agent's *long-term behavioral dispositions*.
Sensory Motor Memory	Stores *motor plan templates* – abstract (partially specified) motor plans that are used to instantiate concrete (fully specified) motor plans for selected behaviors.	Memory of concrete motor plans that allow the instantiation of learned skills and habits. LIDA's motor plans can be thought of as part of an agent's *short-term behavioral dispositions*.
Transient Episodic Memory	Stores recent *episodes* consisting of sequences of events and event-related content, such as portions of perceptual scenes. Episodes that do not decay away are consolidated into Declarative Memory.	Memory of recent (unconsolidated) episodes, which are central to the creation of long-term narratives and are a constituent part of the *narrative self*.
Declarative Memory	Stores autobiographical memories, semantic memories, narratives, and narrative templates (Kronsted et al., 2021).	Memory of (self-)narratives and propositional content that can figure into these narratives. This kind of content is constitutive of the narrative self.

Using this table, we can roughly distinguish between long-term memory modules whose content is relevant to the embodied self and

[8] LIDA's coalitions can be understood as a combination of representational content from LIDA's CSM and one or more attention codelets that advocate for the salience of that content. Two or more attention codelets may *jointly* advocate for the same content in the CSM. Such joint coalitions will typically have additional activation, and thus are more likely to win the competition in LIDA's Global Workspace.

[9] LIDA's schemes were inspired by the "schemas" used in Drescher's schema mechanism (Drescher, 1991).

those relevant to the narrative self. Recall that the narrative self is the aspect of the self-pattern composed of an agent's self-narratives and the processes that create and modify them. Narratives, which we defined as sets of actions and events that are temporally ordered and semantically structured (see section 2), are primarily stored in Declarative Memory. However, narratives will typically include many associations to content in other long-term memory modules, such as PAM. These associations allow narratives to be grounded (Harnad, 1990). The creation and modification of narratives depends on structure building codelets. These are special purpose processes that continually scan LIDA's Workspace, constructing narratives from currently active (e.g., cued) event structures and related content (see Section 5.3.1.). Some of this content may be cued into the CSM from Transient Episodic Memory, which contains an agent's recent episodes, or from Declarative Memory, which contains content such as autobiographical memories. Therefore, an adequate characterization of the narrative self in LIDA will include content from these long-term memory modules, as well as supporting processes like narrative-constructing structure building codelets. The role of Transient Episodic and Declarative Memory in the construction and modification of narratives has been extensively described by Kronsted, Neemeh, Kugele, and Franklin (2021), which we will draw upon in the following sections.

The embodied self, as we have described it in this paper, includes (but is not limited to) those aspects of self that are focused on directing an agent's situated environmental interactions; that is, the "behavioral aspects" of Gallagher's pattern theory of self. The *long-term* behavioral aspects of the self-pattern are most apparent in LIDA's Procedural Memory and Sensory Motor Memory. These long-term memory modules contain representations – schemes and motor plan templates – that can be thought of as encoding behavioral dispositions that are instrumental in the selection and execution of actions. As can be seen in Fig. 3, the production of behaviors in LIDA is a process of continued refinement: First, relevant schemes in Procedural Memory are instantiated, based on the content that was consciously broadcast. These behaviors (instantiated schemes) specify *what can be done*, given the situational context the agent finds themselves in. Many such behaviors may be sent to LIDA's Action Selection module, where they compete for selection.[10] The result of this competition is a single selected behavior that determines *what to do next*. Finally, Sensory Motor Memory instantiates a motor plan for the selected behavior based on an appropriate motor plan template.[11] Motor Plans specify *how* a selected behavior's action can be realized as an adaptable sequence of motor commands. Motor plans are based on subsumption architectures (Brooks, 1986; 1991) that are executed through a process of "online control" (Dong & Franklin, 2015) during which stimuli from the agent's environment (over LIDA's dorsal stream) largely determines the next motor command emitted by the motor plan.

Also part of the embodied self are sensorimotor representations in PAM that relate to embodied aspects of the self such as the agent's body schema and their knowledge about affordances for action. As in the case of the narrative self, there will likely be structure building and attention codelets whose concerns include some of these sensorimotor representations. Therefore, an adequate characterization of the embodied self in LIDA would have to include both the contents described above, as well as the codelets who are concerned with them. An interesting, but open question, is whether there are any codelets whose concerns include contents of both the embodied and the narrative self, therefore providing a kind of "interface" between the two self-aspects. We don't want to take a strong stance on this issue, but we will briefly come back to it near the end of our paper.

The distinction between the different kinds of modules and codelets that constitute the embodied and the narrative self respectively is central to our model of how aspects of the self-pattern interact. In the following section, we will explicate two model cases of this interaction in LIDA, with the aim of clarifying conceptual questions about the mutual interaction of aspects of a self-pattern more generally.

4. Modeling the bi-directional influence of the embodied and narrative self in LIDA

In this section, we will provide a step-by-step description of how the embodied and narrative aspects of an agent's self-pattern may interact during the two example situations we laid out in Section 2. These descriptions will closely follow the processes taking place within individual LIDA modules, such as the CSM, as well as patterns of interaction between different LIDA modules.

4.1. Narrative-to-embodied influence: Self-programming

Our first example is that of an academic named Alice, who comes to the realization that she is neglecting her child. As a result, she decides to change her behavioral routines: Instead of working in her study on Sunday mornings (her accustomed habit), she will take her son Bob to soccer practice. In functional terms, this is an instance of *narrative self-programming*, where the presence of a narrative goal ("I want to be a good mother"), in conjunction with a process of narrative deliberation, leads to persistent changes in behavioral routines over time (Dings, 2018). Moreover, this example illustrates how the narrative self can influence the embodied self.

Our description of narrative self-programming in LIDA focuses on (1) the content and interactions within LIDA's (preconscious) Workspace that initiate narrative deliberation, (2) the influence of self-narratives on LIDA's volitional (deliberative) mode of action selection (Franklin, 2000; Franklin et al., 2016) and (3) the long-term effect of these cognitive processes on the agent's behavioral routines. To understand the interplay between these processes more clearly, let us look at how some of Alice's relevant cognitive cycles may play out as she realized that she has been neglecting her child.

As Alice steps out of her study, she is confronted with the sight of her son, Bob. He is standing in front of her, wearing soccer gear and holding a soccer ball under his arm. Bob is looking at her with a disappointed expression on his face. These environmental stimuli result in the activation of Perceptual Associative Memory (PAM) and the instantiation of relevant percepts into the CSM. These percepts might include perceptual representations about Bob (his appearance, his name, Bob-related feelings) and soccer paraphernalia. After this perceptual content is instantiated into the CSM, it may cue other memories of Bob being disappointed, for instance episodes from Transient Episodic Memory and Declarative Memory. For our example, we assume that Alice's TEM includes a recent episode that Alice might describe as "Earlier today, Bob was disappointed when I told him I couldn't come to his soccer game" (Fig. 4, Left). We also assume that Alice's Declarative Memory contains a narrative goal of wanting to be a good mother.

With these (and more) memory contents active in the CSM, structure building codelets (SBCs) begin looking for content that fits their interests. For instance, a "causality" SBC might build a new structure causally relating Bob's current disappointment and the recent episode in which Alice told Bob she couldn't attend his soccer game (Fig. 4, Right). As a result of this inference, a preconscious representation describing the cause of Bob's disappointment can be encoded in the CSM. A *proposer* (Franklin et al., 2016, Section 6.2) structure building codelet associated with Alice's narrative goal of being a good mother may notice this structure and create a new proposal (i.e., option to act) aimed

[10] Action Selection's behavior selection algorithm is analogous to the winner-take-all competition in LIDA's Global Workspace. (See Franklin et al., 2016 for more details on LIDA's action selection procedure.).

[11] Sensory Motor Memory's instantiation operation (called "specification" by Dong & Franklin, 2015) is analogous to the instantiation operation performed by Procedural Memory that generates "behaviors" from "schemes.".

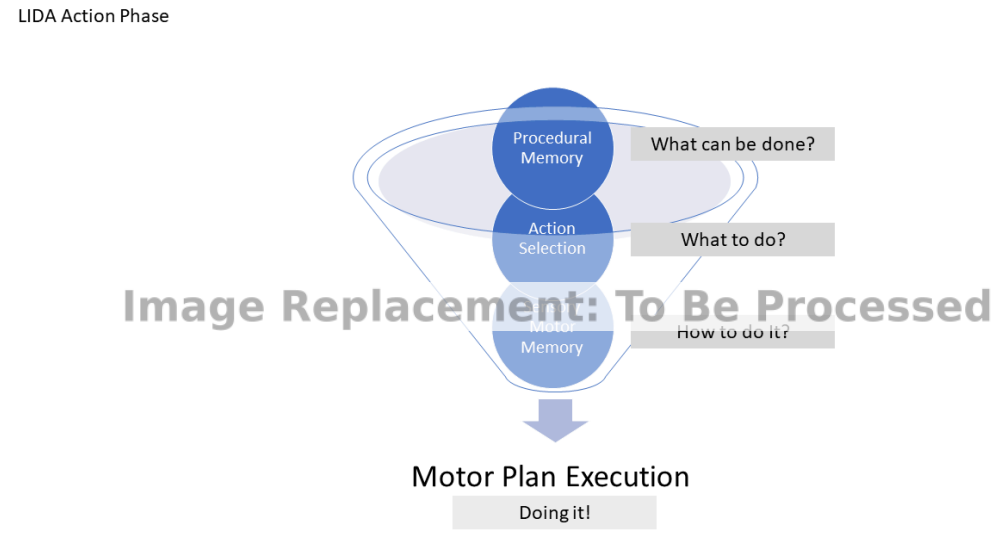

Fig. 3. A high-level schematic of action selection and execution in LIDA.

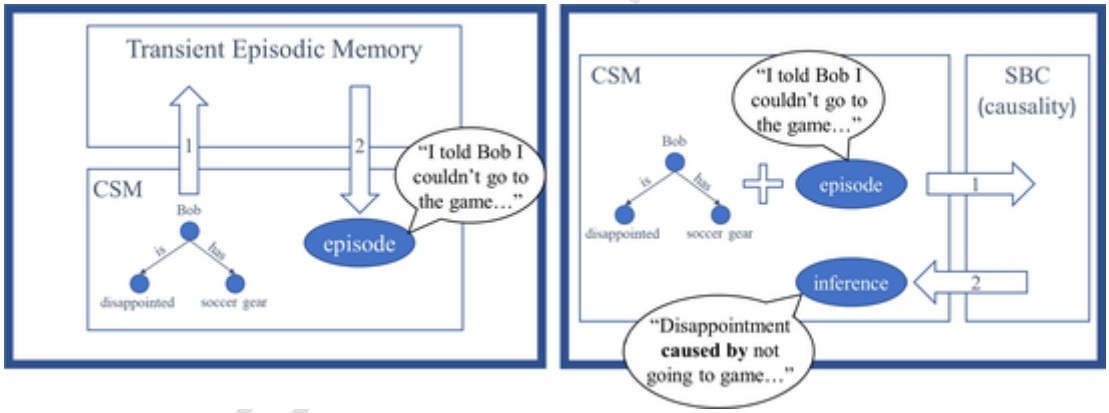

Fig. 4. Modeling the influence of narrative contents and processes on embodied habits in LIDA (Part 1). **Left**: A recent episode is cued from Alice's TEM due to the presence of a structure consisting of nodes representing Bob, his being disappointed, and his soccer gear. **Right**: Structure building codelets looking for causal connections between content in the CSM may then infer the cause of Bob's disappointment based on this structure and the episode.

at alleviating Bob's disappointment: "going to Bob's soccer game" (see Fig. 5, Left).

Concurrently, attention codelets scan the CSM, looking for structures that satisfy their salience criteria. One such attention codelet might look for content with high incentive salience (desirability). Another may look for content with high activation (e.g., situational relevance or strong feelings). For Alice, Bob-related structures may have both high incentive salience (due, in part, to her narrative goal of being a good mother) and activation (due to her current encounter with Bob), so these structures are statistically more likely to become part of a winning coalition. Therefore, the coalitions containing Bob-related content will likely have very high activations. Consequently, a coalition containing the proposal to attend Bob's soccer game wins the competition for consciousness and is globally broadcast (Fig. 5, Right, Steps 1 & 2). Procedural Memory receives this conscious broadcast, which it uses to activate its schemes (Fig. 5, Right, Step 3). Among these is a "go to {LOCATION}" scheme,[12] which is relevant to the current proposal. As a result, this scheme is instantiated as a "go to soccer game" behavior (i.e., "soccer game" is bound to the "{LOCATION}" variable) and sent to Action Selection (Fig. 5, Right, Step 4). Other schemes may also be instantiated and sent to Action Selection if their contexts or results are relevant to this conscious broadcast.

These behaviors then compete for selection in LIDA's Action Selection module. In this case, a previously selected "work" behavior remains in Action Selection from an earlier cognitive cycle. (Recall that Alice just left her study when she encountered her son Bob.) Given Alice's accustomed habit of working during this time, and the fact that she was recently engaged in work-related activities, the current "go to soccer game" proposal is too close in activation to the "work" behavior to win the competition in Action Selection outright. Instead, a special *deliberation behavior* is selected, which initiates LIDA's volitional mode of action selection.[13] As subsequent cognitive cycles unfold, Alice deliberates about whether to "go to the soccer game" (Fig. 5, Right, Step 5).

[12] {LOCATION} is an unbound variable, which is given a specific value during instantiation. Using variables, schemes can be generalized to operate in many different situational contexts.

[13] LIDA has four modes of action selection: consciously mediated, volitional (e.g., deliberative), automatized, and alarms. LIDA's implementation of volitional decision making (Franklin, 2000; Franklin et al., 2016, sec. 6.2) is based on James's Ideomotor Theory (James, 1890).

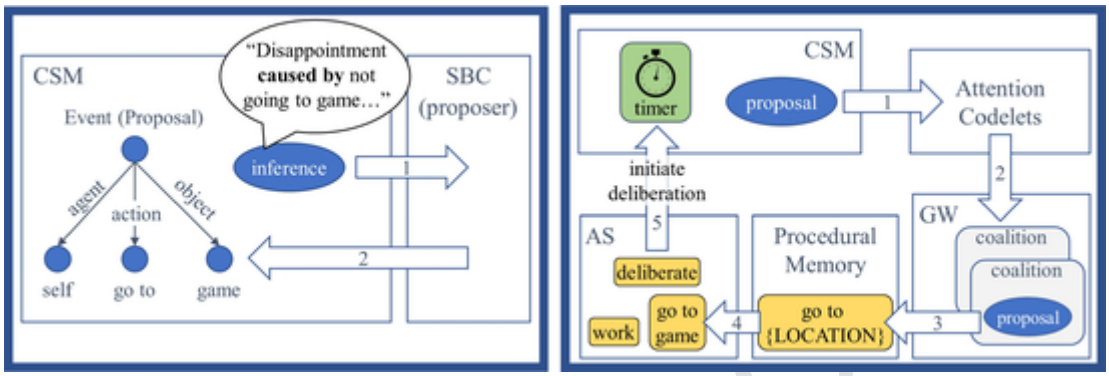

Fig. 5. Modeling the influence of narrative contents and processes on embodied habits in LIDA (Part 2). **Left**: A structure building codelet creates a new *proposal* based on the previously created inference. **Right**: This proposal is globally broadcast, resulting in the instantiation of a "go to [soccer] game" behavior (i.e., "game" is bound to the scheme's "LOCATION" variable). This behavior competes for selection in Action Selection (AS) with other behaviors, such as a "work" behavior. In this case, a deliberation behavior is selected because no selectable behaviors have sufficient activation to win outright. At the beginning of deliberation, a "timer" is started in the CSM that governs the deliberation process. If the timer expires without "objections" to the proposal, then the proposal is accepted and converted into a goal.

Already at this point, we can see that the LIDA model incorporates a lot of mechanisms that can be conceptualized as long-term *cognitive dispositions*: For instance, Alice's narrative goal of wanting to be a good mother may not just be realized by the presence of semantic content in declarative memory (such as "I want to be a good mother"), but also by changes in the base-level activation of structure building and attention codelets.[14] Another way that such a goal may be realized is by the creation of *new* structure building and attention codelets, which look out for content relevant to the narrative goal in particular. This way of modeling narrative goals is in accord with Slors' conceptualization of distal intentions as cognitive dispositions (Slors, 2015). For now, let us return to the case of Alice and Bob.

One of the structures present in Alice's CSM is a proposal to take Bob to his soccer game, which was created by a Structure Building Codelet building proposals for solving conflicts. If this proposal – as part of a coalition – wins the competition and is broadcast to consciousness, Alice comes to the realization that her son is disappointed because he wants to go to the soccer game with her. It also makes her realize that she can do something about this situation by going to the soccer game with him. Furthermore, the proposal instantiates a scheme from Procedural Memory that is related to the event it consists of – namely going to the soccer game with Bob.[15] Within Action Selection, this instantiated scheme (i.e. behavior) then competes with other behaviors that have recently become instantiated (such as a behavior for working in her office). If there is no clear winner of this competition, a special *deliberation behavior* wins the competition instead, allowing the agent to consciously deliberate the options available to them (Fig. 5). In his case, both the "work" and "take Bob to soccer" behaviors have a similar level of activation, so that neither of them is immediately selected. Therefore the deliberation behavior wins the competition, and a new cognitive cycle unfolds, during which the agent is in a mode of deliberation.

During deliberation, structure building codelets function as *objectors* or *supporters* (Franklin et al., 2016, Section 6.2) for the current proposal. In general, any structure in the CSM that is situationally relevant to this proposal can function as an "objection" or "support" for the proposal, depending on its content.[16] In Alice's case, two strongly relevant factors for her decision have to do with her self-narratives: On the one hand, Alice sees herself as a hard worker who always does her best to get her work handed in on time. But on the other hand, Alice also has the narrative goal of wanting to be a loving mother, and believes that loving mothers should not disappoint their children. This narrative dilemma can be modeled in LIDA by a series of deliberative cognitive cycles during which objectors and supporters build structures from narrative content cued from Semantic Memory. For instance, just after deliberation starts, an objector SBC might construct an objection based on Alice's self-narrative of being a hard worker, thus preventing the immediate acceptance of the proposal (Fig. 6).[17] However, during the next cognitive cycle, another SBC might build a support structure, based on her narrative goal of wanting to be a good mother and a narrative template about how "good mothers" treat their children (e.g., not disappointing them). This additional support for the proposal resumes Alice's process of narrative deliberation. Assuming no further objections, the deliberated proposal (taking Bob to his soccer game) is accepted.[18] Thus, Alice's narrative deliberation ends with her making a decision to take Bob to his soccer game instead of going back to work in her study. This results in the formation of a new distal intention (Kronsted et al., 2021), which can later support planning and the production of goal-directed behaviors, such as finding the keys to Alice's car.

The architectural components that implement Alice's distal intentions and narrative goals (e.g., "I want to be a good mother") may extend well beyond semantic content in Declarative Memory. New structure building codelets might be spawned that build preconscious representations specific to a narrative goal. This content may, for example, aid an agent's goal-specific situational understanding, or specify preconscious "options" (see Franklin et al., 2016, Section 6.2) that function as "sub-goals" towards the attainment of a narrative goal. Similarly, attention codelets can be spawned that bias an agent's attention towards goal-relevant content in the CSM. Collectively, these many mechanisms can be viewed as components of long-term *cognitive dispositions*, that exert a persistent influence on an agent's behaviors.

[14] Attention and Structure Building Codelets both have a base activation that factors into the level of activation assigned to the structure/coalition that they build. For the purpose of simplification, we will not talk about this part of the activation-determining process, but it's worth noting that this dispositional feature of activation distribution is found at various steps within the cognitive cycle. For more information, see Franklin et al. 2016.

[15] More accurately, since this is an action-only scheme, it's going to be less specific, e.g., consisting of preparatory movements to going *somewhere* by car.

[16] Both in our general characterization and in LIDA, deliberation does not necessarily involve only propositional content. In our current example, we nevertheless focus on narrative deliberation, which is propositional in nature.

[17] Computationally, this "stops" the deliberation timer. Without additional support, the deliberation process will end with the current proposal being rejected.

[18] As seen in Fig. 6, the objector/supporter implementation in LIDA typically involves a timer codelet, which we omit from our description for the sake of scope. For more details about action deliberation in LIDA, see Franklin (2000).

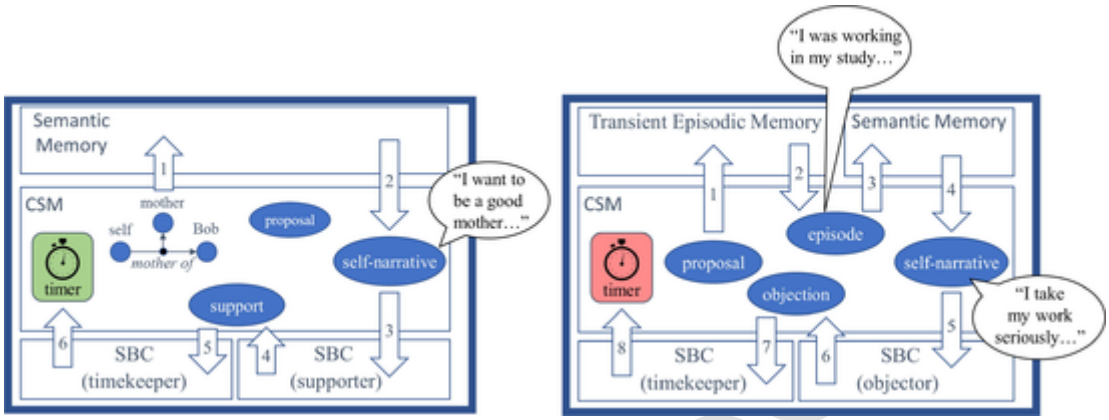

Fig. 6. Modeling the influence of narrative contents and processes on embodied habits in LIDA (Part 3). **Left**: Content in the CSM (e.g., a recent episode of Alice working in her study) cues Alice's self-narrative of being a hard worker. An objector (SBC) recognizes this self-narrative as conflicting with the current proposal, and generates an objection. This objection stops the timer before it expires, preventing the proposal from being immediately accepted. **Right**: In a similar manner, content in the CSM defining the relation between Alice (mother) and Bob (son) cues a self-narrative about Alice wanting to be a good mother. A supporter (SBC) uses this self-narrative to build a "support" structure for the current proposal, and the deliberation timer is restarted. The proposal is accepted if no further objections occur before the timer expires (ending deliberation).

Later in the day, when Alice talks with her son about how his day was, the entire event of her making the decision, taking her son to the soccer game, and experiencing his joyful reaction may enter into her CSM. From there, it can cue other related memory contents, for instance Alice's narrative about wanting to be a good mother. These preconscious dynamics, coupled with the positive feelings that Alice begins to associate with taking Bob to his soccer game, may give rise to new behaviors such as promising Bob to take him to his game every Sunday from now on. This, in turn, may give rise to new self-narratives and narrative goals that further influence Alice's behaviors (e.g., "Every Sunday I take Bob to his soccer game"). As autonomous agents are often prone to continuing with an established habit instead of changing it, it is likely that the new scheme "go to soccer practice with Bob" might initially have a lower base-level activation than her habitual scheme "go into your study and work".

There are, however, a few additional factors that can support Alice's distal intention to take her son to his soccer game every Sunday. For instance, along with the new scheme, new structures will be created in Alice's PAM that associate Sunday mornings with soccer equipment and her son. Additionally, Alice may make associations between episodes where she took Bob to his soccer game and the positive feeling of seeing him enjoying himself, which may make her distal intention to take him there next week as well more desirable and therefore more likely to be realized. Over time, these factors, combined with an increasing number of events in which Alice actually takes Bob to soccer practice, will likely increase the base-level activations of the schemes associated with these activities, and the incentive salience of those outcomes to the point where these schemes become new habits for Alice.[19] Eventually, Alice may abandon an old behavioral disposition (going to her study) in favor of a new one (taking her son to the soccer game) – this is exactly the process of "overriding" an established behavior we talked about in section 2. Thus, we now have a case where the distal intention that was created through a momentary process of narrative deliberation, results in long-term changes to the embodied self (Alice's behaviors and situational associations) via the formation of a narrative goal as a distal intention, following a period of gradual reinforcement.

4.2. Embodied-to-narrative influence: Modification of narratives

Now consider the case of a non-sportive person who increasingly identifies as a tennis player due to the adoption of a new behavioral routine (see Section 3.2). In LIDA, behavioral routines are modeled as *behavior streams*: higher-order behaviors that contain *composite actions*.[20] Using composite actions, schemes in Procedural Memory can be nested, where the coordinated selection and execution of lower-level behaviors (e.g., joint movements) can constitute a higher-level action (e.g., a tennis forehand). For instance, once our tennis player has accompanied her friend to tennis practice a few times, she might form a behavior stream for a set of activities (behaviors) that lead to her playing tennis with her friend (meeting up with her friend, taking the bus to the tennis court, changing into her tennis gear, etc.).[21] Each of these component behaviors may, themselves, be implemented as a behavior stream.

In general, behavioral routines develop gradually. Through environmental interactions, agents learn new schemes, enriching their available behavioral repertoire. Some of these schemes can be instrumental towards the achievement of a composite action's goal state.[22] By repeatedly engaging in an activity, individual schemes relevant to that activity can be reinforced, and implicit chains of schemes can be identified in Procedural Memory. That is, if a scheme reliably results in an outcome that satisfied the context of another scheme, LIDA's Procedural Memory module can learn to associate these scheme with each other. It is the identification of these implicit chains of schemes that

[19] Habits (i.e., behavioral routines) are frequently executed behavior streams that are likely to be selected by Action Selection unless the agent faces situational challenges that either cause other behaviors to be more necessary to execute or the habit's context to not be sufficiently fulfilled. Behavioral routines are discussed in more detail in section 4.2.

[20] Composite actions are implemented using *chains of schemes*. A chain of schemes is a sequence of schemes in which the context of each scheme (after the first) is satisfied by the result of the scheme that precedes it.

[21] Behavior streams often consist of many different chains of behaviors that bring about the same result, since there may be many stable ways to achieve the same outcome using different actions. However, for our purposes, it will be sufficient to assume the simplest case in which a stream consists of only one chain of schemes.

[22] A composite action's goal state is typically different from the "result" of a scheme. For example, a composite action may contain an adaptable set of bodily movements that affect the toggling of a wall switch to its up position. The goal state of this composite action would be the switch in the up position. The result of a scheme containing that composite action, on the other hand, may specify that a lamp is turned on or off, or that a garbage disposal turns on or off, or that nothing happens at all. Like all actions (primitive and composite) the results depend heavily on the scheme's context (e.g., whether the agent is in the bedroom or kitchen when executing its action).

eventually gives rise to the higher-order behaviors that specify behavior streams and behavioral routines.

This learning process is greatly facilitated by the occurrence of schemes with desirable results, which incentivize agents to engage in an activity. In terms of LIDA, a behavior (instantiated scheme) with a desirable result is more likely to be selected by Action Selection, and desirable content is more likely to be globally broadcast. In our case of the initially non-sportive person accompanying her friend to the tennis court, these desires may start as positive feelings of fun, excitement or accomplishment induced by successfully hitting a ball across the net. These positive feelings increase the incentive saliences (desirability) of those individual events, and, more generally, the desirability of playing tennis. These desires can become associated with the content comprising an agent's schemes, eventually leading to behavior streams that effect positively valued environmental states. Thus, as she keeps accompanying her friend to the tennis court, her schemes that specify the lower-level actions necessary for engaging in this behavior can be learned and reinforced, and give way to behavior streams. In this way, behaviors in LIDA can become routine, constituting contextually anchored dispositions to act.

The presence of a widely applicable behavioral routine makes it more likely for the agent to frequently encounter the same environmental situations – namely those that fit the results specified by its underlying behavior stream. More generally behavioral dispositions can cause an agent to become increasingly acquainted with the environmental conditions that allow for, and result from, the actualization of that disposition. For instance, once our initially non-sportive person has formed the behavioral routine of going to the tennis court with her friend, she will regularly be exposed to tennis-related sensory input. Such regular exposure to the same environmental stimuli will naturally lead to the creation of new associations between those stimuli and the perceptual and semantic representations that serve to conceptualize its sensory landscape. For instance, through repeated exposure to play, our tennis player will eventually be able to distinguish forehand from backhand strokes, aggressive from defensive moves, and so on. Furthermore, she – or, more precisely, her structure building codelets – will be able to construct (self-)narratives and narrative templates based on tennis-related event structures and propositions about such things. This can also give rise to narrative goals. All of these are narrative aspects of the self-pattern.

Continuing with our example, suppose our tennis player's friend is no longer able to join her at practice after a while. Nonetheless, she keeps going to practice, and eventually, her friend asks her, "I've noticed you're still going to practice! Are you becoming a serious tennis player now?" This causes the creation and conscious broadcast of a proposal ("option") to answer this in a positive way: "Yes I am!" Moreover, due to her adoption of a new tennis-related behavioral routine and her increasing desire to win tennis matches, a behavior instantiated for this proposal is likely to be accepted (by Action Selection) following a brief moment of deliberative self-reflection. Later, this episode in which she affirms her identity as a serious tennis player, may be cued into the CSM and used by a structure building codelet to create a new self-narrative about her being a tennis player. This new self-narrative will then be reinforced whenever it comes to consciousness, which is likely to happen more often as she spends more of her free time playing tennis. In contrast, her old self-narrative of being a non-sportive person is increasingly unlikely to come to consciousness, as it no longer matches the concerns of the agent's structure building and attention codelets. As a result, that old self-narrative will likely decay away and eventually be forgotten. Thus, if the agent must make a decision in which her self-narrative is a relevant factor – for instance, deciding whether to pursue an international tennis career or stay in her hometown with a stable job – this decision will be influenced strongly by her new self-narratives, which were ultimately brought into existence by her adoption of a new behavioral routine. In this way, the establishment of a behavioral routine – a component of the embodied self – can lead to long-term adjustments to the narrative self (e.g., the replacement of old self-narratives with new ones).

5. Practical implications

The manner in which self-patterns arise, are maintained, and change within agents has a profound impact on how these agents are able to learn about and interact with their environments. Human agents, for instance, are able to modify their environments both in the short and in the long term in order to facilitate the acquisition of knowledge and skills that allow them to reach their self-defined goals (Knowles, 1975; Gureckis & Markant, 2012). This ability is based, at least in part, on the agent's current knowledge about themselves, particularly about their short- and long term goals, as well as about their current level of expertise. This kind of self-knowledge is becoming increasingly relevant in the field of Artificial General Intelligence (AGI) as well, where there have been recent attempts to build artificial agents capable of *self-directed learning*, a type of learning that requires the agent to be in voluntary control of the conditions under which episodes of learning and skill acquisition occur (Kumar, Singh, & Buyya, 2021; Dannenauer, 2021). In artificial agents, self-directed learning can be seen as a type of self-supervised learning that takes into account agent-internal intentional states, such as the agent's knowledge about its own learning process and the goals it establishes and reaches during learning. A recent attempt at formalizing the application of self-directed learning in artificial agents has been made by Zhu, Wang, and Xie (2022), who identify self-awareness as a key factor in enabling self-directed learning. They note that human learners typically become familiar with their own learning paces and styles over time, allowing them to choose learning strategies that benefit them the most within a given task environment. Similarly, an artificial agent that is aware of its own learning patterns will show more flexibility and greater learning speed when faced with different and/or changing task environments. In the context of motor learning, conceiving of self-related knowledge in terms of the pattern theory of self has some clear advantages in comparison to the classic conceptualization of the self as a unitary entity: Changes to processes of an agent's embodied self (such as during interaction learning) do not directly lead to the modification of other, already-established processes within the agent. Rather, there is a temporal dynamic between processes associated with the different self-patterns (e.g., those of the embodied and the narrative self) that determine the degree to which, for instance, the learning of a new embodied skill affects the agent's views about themselves and their actions within a larger context. In LIDA, this temporal dynamic is realized by a combination of the model's structural features (which processes and modules exist and how they interact) and the temporal succession of cognitive cycles (over which timescales cognitive processes play out). However, in theory, insights from the pattern theory of self should be applicable to many other AGI architectures: Abandoning the commitment to the existence of a unitary, stable self-entity can benefit the agent's autonomy, adaptivity and learning speed regardless of which particular architecture the pattern theory is implemented in.

Specific use cases for the pattern theory of self include the development of AGI for general applications of human–machine interaction, such as customer support, workflow optimization, policy development, or medical diagnosis (Strain, Kugele, & Franklin, 2014). In all of these cases, we contend that human–machine interaction is most fruitful and successful if the AGI in question possesses both a sense of self and a sense of other similar to those found in humans. This may allow autonomous agents to both understand their user's needs more closely and actively coordinate its own learning process with that of its users. These benefits make an implementation of a multifaceted sense of self an important goal for any AGI architecture.

Finally, the temporal dynamic between self-patterns modeled here may shed light on another issue relevant to AGI research, namely the interface problem: How do perceptual representations, which are often assumed to be non-semantic, interface with conceptual representations which bear semantic content? Our distinction between different temporal scales characteristic for certain cognitive processes may provide a solution for this problem as well: Perceptual and conceptual representations can become involved with both short-term and long-term cognitive processes. In the case of dispositions of the embodied and the narrative self (and vice-versa), we saw that the influence of one self-aspect on another happens across multiple temporal scales: For instance, Alice underwent a moment of narrative deliberation, from which the selection of a certain behavior emerged, which in turn became stabilized as a persistent behavioral routine as Alice experienced positive feedback following her decision. In a similar sense, a possible solution to the interface problem may take into account types of agent-internal processes that play out across different temporal scales, therefore eliminating the need for an "interface" to exist at a concrete step along a single cognitive process at all. Taking inspiration from LIDA, a possible alternative conception would then be a replacement of the "interface" with that of a "workspace", which representations of any format may inhabit, and from which an agent's beliefs, goals, memories, and narratives may be built over the course of various numbers of cognitive cycles. Alternatively, structure building codelets in the Workspace could be conceived as "interfaces" between different types of representations, by building structures that include both non-semantic perceptual and semantic conceptual representations. Which, if any, of these two solutions is more conceptually sound may become the topic of future modeling efforts in LIDA.

6. Conclusion

In our paper, we provided a functional implementation of the pattern theory of self within the LIDA model. We focused in particular on implementing the reciprocal interaction between aspects of a self-pattern, using embodied and narrative aspects as exemplar cases. In principle, our findings should be generalizable to other aspects of self-patterns as well, including for instance affective, intersubjective, or normative aspects (Gallagher and Daly, 2018). Similar to how the embodied and the narrative self interact with each other over both short and long time scales, other aspects of an agent's self-pattern will also interact across multiple temporal scales, producing stable but adaptive dispositions to think or act in certain ways. In order to understand the decision-making process of both human and artificial intelligent agents in more detail, it will be important to take these different temporal scales into account. We have shown that the LIDA model can functionally account for these temporal differences when it comes to the processes establishing and maintaining a self-pattern. The next step will be to build an agent that incorporates this and more insights from the pattern theory of self (Ryan, Agrawal, & Franklin, 2020), which may finally bring us closer to building artificial agents capable of maintaining a sense of self that can influence their decision-making in the same way as it does in human agents.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Albert Newen's work is supported by the NRW Profillinie "Interact!" (PROFILNRW-2020-135).

Data availability

No data was used for the research described in the article.

References

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Baars, B.J., Franklin, S., & Ramsøy, T.Z. (2013). *Global workspace dynamics: Cortical "binding and propagation" enables conscious contents. Frontiers in Psychology*, *4*, 200. https://doi.org/10.3389/fpsyg.2021.749868.

Baars, B.J., Geld, N., & Kozma, R. (2021). *Global workspace theory (GWT) and prefrontal cortex: Recent developments. Frontiers in Psychology*, *5163*. https://doi.org/10.3389/fpsyg.2013.00200.

Brooks, R. (1986). *A robust layered control system for a mobile robot. IEEE journal on robotics and automation*, *2*(1), 14–23.

Brooks, R. A. (1991). *Intelligence without representation. Artificial intelligence*, *47*(1–3), 139–159.

Cox, M.T. (2005). *Metacognition in computation: A selected research review. Artificial Intelligence*, *169*, 104–141. https://doi.org/10.1016/j.artint.2005.10.009.

Cutsuridis, V., Hussain, A., & Taylor, J.G. (2011). *Perception-Action Cycle. Models, Architectures, and Hardware: Vol. 1*. Springer.

Dijkstra, T.M.H., Schöner, G., & Gielen, C. (1994). *Temporal stability of the action-perception cycle for postural control in a moving visual environment. Experimental Brain Research*, *97*, 477–486.

Dings, R. (2018). *The dynamic and recursive interplay of embodiment and narrative identity. Philosophical Psychology*, *32*(2), 186–210. https://doi.org/10.1080/09515089.2018.1548698.

Dings, R. (2021). *Meaningful affordances. Synthese*, *199*, 1855–1875. https://doi.org/10.1007/s11229-020-02864-0.

Dong, D., & Franklin, S. (2015). *A New Action Execution Module for the Learning Intelligent Distribution Agent (LIDA): The Sensory Motor System. Cognitive Computation*. https://doi.org/10.1007/s12559-015-9322-3.

Drescher, G.L. (1991). *Made-up minds: A constructivist approach to artificial intelligence*. MIT Press.

Franklin, S. (2000). *Deliberation and Voluntary Action in "Conscious" Software Agents. Neural Network World*, *10*, 505–521.

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

Dannenauer, D. et al. (2021). Self-directed Learning of Action Models using Exploratory Planning. Proceedings of the 9th Annual Conference on Advances in Cognitive Systems, 1–18. 10.48550/arXiv.2203.03485.

Franklin, S. & Baars, B. (2010). Two varieties of Unconscious Processes. In: E. Perry, D. Collerton, H. Ashton & F. LeBeau (Eds.): *New Horizons in the Neuroscience of Consciousness*, John Benjamin.

Franklin, S. & Grasser, A. (1997). Is It an agent, or just a program? A taxonomy for autonomous agents. In: Müller, J. P., Wooldridge, M. J., Jennings, N. R. (Eds.): *Intelligent Agents III: Agent Theories, Architectures, and Languages*. ATAL 1996. Lecture Notes in Computer Science, Vol. 1193. Springer. 10.1007/BFb0013570.

Fuster, J. M. (2002). Physiology of executive functions: The perception-action cycle. In D. T. Stuss & R. T. Knight (Eds.): *Principles of frontal lobe function* (pp. 96–108). Oxford University Press. 10.1093/acprof:oso/9780195134971.003.0006.

Fuster, J.M. (2004). *Upper processing stages of the perception-action cycle. Trends in Cognitive Sciences*, *8*(4), 143–145. https://doi.org/10.1016/j.tics.2004.02.004.

Gallagher, S. (2013). *A pattern theory of self. Frontiers in Human Neuroscience*, *7*(443). https://doi.org/10.3389/fnhum.2013.00443.

Gallagher, S., & Daly, A. (2018). *Dynamical Relations in the Self-Pattern. Frontiers in Psychology*, *9*(664). https://doi.org/10.3389/fpsyg.2018.00664.

Gureckis, T.M., & Markant, D.B. (2012). *Self-Directed Learning: A Cognitive and Computational Perspective. Perspectives on Psychological Science*, *7*(5), 464–481. https://doi.org/10.1177/1745691612454304.

Goertzel, B., & Pennachin, C. (2007). The Novamente Artificial Intelligence Engine. In Goertzel, & Pennachin (Eds.), *Artificial General Intelligence*. Springer.

Harnad, S. (1990). *The symbol grounding problem. Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346.

James, W. (1890). *The principles of psychology*. Cambridge, MA: Harvard University Press.

Knowles, M.S. (1975). *Self-Directed Learning: A Guide for Learners and Teachers*. Association Press.

Kotseruba, I., & Tsotsos, J.K. (2018). *40 years of cognitive architectures: Core cognitive abilities and practical applications. Artificial Intelligence Review*, 1–78. https://doi.org/10.48550/arXiv.1610.08602.

Kronsted, C., Neemeh, Z.A., Kugele, S., & Franklin, S. (2021). *Modeling long-term intentions and narratives in autonomous agents. Journal of Artificial Intelligence and Consciousness*, *8*(1), 229–265. https://doi.org/10.1142/S2705078521500107.

Kugele, S., & Franklin, S. (2021). *Learning in LIDA. Cognitive Systems Research*, *66*, 176–200. https://doi.org/10.1016/j.cogsys.2020.11.001.

Kumar, J., Singh, A.K., & Buyya, R. (2021). *Self directed learning based workload forecasting model for cloud resource management. Information Sciences*, *543*, 345–366. https://doi.org/10.1016/j.ins.2020.07.012.

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

McCall, R., Franklin, S., Faghihi, U., Snaider, J., & Kugele, S. (2020). *Artificial motivation for cognitive software agents. Journal of Artificial General Intelligence*, *11*(1), 38–69. https://doi.org/10.2478/jagi-2020-0002.

Morris, T.H. (2019). *Self-directed learning: A fundamental competence in a rapidly changing world. International Review of Education*, *65*, 633–653. https://doi.org/10.1007/s11159-019-09793-2.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. W. H. Freeman.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
Newen, A. (2018). *The Embodied Self, the Pattern Theory of Self, and the Predictive Mind*. *Frontiers in Psychology*, *9*(2270). https://doi.org/10.3389/fpsyg.2018.02270.
Newen, A., & Vogeley, K. (2003). *Self-Representation: Searching for a neural signature of self-consciousness*. *Consciousness and Cognition*, *12*(4), 529–543. https://doi.org/10.1016/s1053-8100(03)00080-1.
Ryan, K., Agrawal, P., & Franklin, S. (2020). *The pattern theory of self in artificial general intelligence: A theoretical framework for modeling self in biologically inspired cognitive architectures*. *Cognitive Systems Research*, *62*, 44–56. https://doi.org/10.1016/j.cogsys.2019.09.018.
Slors, M. (2015). *Conscious intending as self-programming*. *Philosophical Psychology*, *28*(1), 94–113. https://doi.org/10.1080/09515089.2013.803922.
Snaider, J., McCall, R., & Franklin, S. (2012). *Time production and representation in a conceptual and computational cognitive model*. *Cognitive Systems Research*, *13*(1), 59–71. https://doi.org/10.1016/j.cogsys.2010.10.004.
Kugele, S. & Franklin, S. (2020). A study in activation: Towards a common lexicon and functional taxonomy in cognitive architectures. Proceedings of the 18th Annual Meeting of the International Conference on Cognitive Modeling, 138–144.
Strain, S., Kugele, S., & Franklin, S. (2014). The learning intelligent distribution agent (LIDA) and medical agent X (MAX): Computational intelligence for medical diagnosis. *2014 IEEE Symposium on Computational Intelligence for Human-like Intelligence (CIHLI)*, pp. 1-8 10.1109/CIHLI.2014.7013390.
Zhu, W., Wang, X., & Xie, P. (2022). *Self-directed machine learning*. *AI Open*, *3*, 58–70. https://doi.org/10.1016/j.aiopen.2022.06.001.