Self-awareness survey

Mohammad Rahmani

1 Surveys

2 Definitions

Generative models Generative models facilitate predicting future states

Descriptive models enable the selection of the representation that best fits the current observation.

Self-awareness **Literal definition** implies that Self-awareness must first rely on perception of self as different from the environment and from other agents Chatila et al. (2018).

Def 1 The capacity to become the object of one's own attention, which arises when an agent focuses not only on the external environment but also on the internal milieu. The agent becomes a reflective observer, processing self-information. It becomes aware that it is awake and actually experiencing specific mental events, emitting behaviors, and possessing unique characteristics Regazzoni (2020)[1].

Def 2 - private vs public Regazzoni (2020)[2]

- **Private** Private self-aspects relate to externally **unobservable** events and characteristics such as emotions, physiological sensations, perceptions, values, goals, and motives
- **Public** self-aspects are visible attributes such as behavior and physical appearance

Approaches to tackle SA Common aspects of the proposed approaches lie on the conception of SA as

- a cognitive embodied process composed of representational and inferential operations of an agent situated in an environment,

- an agent's property which emerges in various forms, including the extent of the SA capabilities ("levels") [1], [6] and the scope of the processed information ("private and public") [2], [7]

Application of SA More recently, SA concepts have been transferred to artificial systems aiming at either
designing intelligent agents or *analyzing their behavior*.
Why SA in AI? The driving motivation for the transfer of biological SA concepts to artificial systems is to improve **autonomy**, **robustness**, and scalability and has been investigated in different fields, including software engineering, machine learning, and robotics [8], [9], [10], [11], [12], [13], [14], [15].
Challenge: A fundamental challenge in most of these approaches is how to systematically integrate SA capabilities into artificial agents.

SA in computational context In a computational context, self-awareness (SA) is a capability of an autonomous system to describe the acquired experience about itself and its surrounding environment with appropriate models and correlate them incrementally with the currently perceived situation to expand its knowledge continuously.

Definition from sensor data and signal processing perspective : An artificial agent is considered self-aware if it can dynamically observe itself and its surrounding environment through different **proprioceptive** and **exteroceptive** sensors and **learn** and **maintain a contextual representation** by processing the observed multi-sensorial data.

Prospective vs exteroceptive : Proprioceptive sensors measure the internal agent's parameters, whereas exteroceptive sensors observe the agent's environment (cp. Regazzoni (2020) Fig 1).

SA introspection The SA representation obtained by jointly and dynamically analyzing the sensory data endows the agent with introspection at different hierarchical levels.

Introspection associates with the agent's capability of estimating and representing *dynamical causal relationships* from the observed sensory data. Such representation allows the agent to model interactions between itself, as observed through proprioceptive sensors; and the environment, as observed through exteroceptive sensors.

importance of embodying SA capabilities The extent of the embodied SA capabilities influences the agent's performance when solving tasks and are assumed as reasons for the significant capability differences of the various biological species.

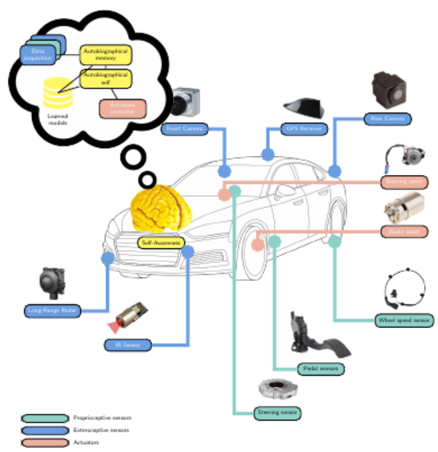

Fig. 1: Concept of a physical architecture for a self-aware autonomous system. The self-aware agent (here conceptually embedded in a vehicle) observes its surrounding environment with exteroceptive sensors (blue) and its internal state with proprioceptive sensors (green) and translates its autonomous decisions into actions through the actuators (in red). The SA core (yellow) is established based on internal representations from the autobiographical memory and the autobiographical self, together with a set of already learned models. The SA core is able to forecast the next state of the environment and the system itself, detects anomalies and executes the derived actions.

Figure 1: Regazzoni (2020)

Minimum requirements to consider an agent self-aware The following capabilities as the minimum requirements in to consider an agent self-aware:

- initialization
- inference
- anomaly detection
- model creation
- interface with control

Regazzoni (2020) Table I describes the proposed SA capabilities and provides a relationship between each of them and biological agents, demonstrating how humans address these capabilities.

3 bio-inspired self-awareness theories

3.1 Damasio's model

autobiographical memories (AMs) Neuroscientists such as Damasio Regazzoni (2020)[28] have provided evidence that neural patterns in the oldest parts of the human brain are organized to process and combine proprioceptive and exteroceptive sensorial information according to hierarchical neural layouts culminating into so-called *autobiographical memories (AMs)*.
AMs can constitute a sort of database for memorizing models of **episodes** that

TABLE I: Definition of self-awareness capabilities and biological relationships.

Self-awareness capability	**Definition**	**Biological relationship**
Initialization	It refers to the initial knowledge from which an agent starts building its own memories. Such initial knowledge provides the agent with the essential tools to interact with its surroundings.	The basic structure of the brain is laid down primarily during the prenatal period, where its *initialization* depends largely on genetics [16].
Memorization	It refers to the agent's capacity of storing and retaining information such that it can be recovered and exploited in the future.	Long-term memories are stored throughout the brain as groups of neurons that fire together in the same pattern that created the original experience. Such operation is done by the process of memory allocation [17].
Inference	It consists of the agent's ability to make predictions about its own future states and its surroundings depending on its current state.	The brain is responsible for anticipating future events. The *predictive coding* theory [18] states that at each level of a cognitive process, the brain generates beliefs of the information it should be receiving from the level below it. These beliefs are translated into predictions about what should be experienced in a given situation.
Anomaly detection	It consists of the agent's ability to recognize observations that cannot be explained by its memories. These observations represent new events that the agent has not detected so far.	Brain predictions are sent as feedback to low-level sensory regions of the brain. The brain then compares its predictions [19] with the actual received sensory input and "explains" high differences (prediction errors) between them.
Model creation	It refers to the agent's capability of generating models that encode previous experiences, facilitating the prediction of the agent's future states and the posterior comparison with evidence.	The prediction errors that can't be explained away get passed up through connections to high levels of feedforward signals, where they are considered newsworthy. The internal models get adjusted so that the predicting error gets suppressed [20].
Decision-making influence	It refers to the ability to generate signals that can be employed by the agent's control system such that its actions are self-monitored dynamically	Muscles move based on commands from the brain [21]. Nerve cells in the spinal cord, called motor neurons, enable to convey and evaluate the brain's commands to the muscles.

Figure 2: Table I

the agent has learned from previous experiences Regazzoni (2020)[5].
Bio-inspired AMs have already been investigated towards implementing self-awareness in artificial agents, for example, in Regazzoni (2020)[29]. Based on anatomical observations, Damasio suggests that episodes in AMs are **represented by a language coding proprioceptive and exteroceptive information according to a temporally ordered causal representation.** Fig. 2 depicts the combination of estimations of the agent's own the external world's state obtained by an early neural layout (named "proto" and "core", respectively) in the form of temporal-causal AM patterns.

According to Damasio, AM patterns are based on *first-person situational descriptors* that enable human agents to represent experienced episodes on the basis of a **neural vocabulary** (i.e., information units). These descriptors always represent exteroceptive data as contextualized to information coming from the agent's body, and vice versa. Thus, patters encoding episodic experiences are represented by coupling the agent and its dynamic interaction with the surrounding environment.
Elementary information units used in AMs define a temporal representation where an agent and the environment reciprocally take on the role of a *context.* Temporal changes of the internal representation of the state of one of them (that Damasio calls "dispositions") are observed as occurring in the context of the other one assuming a given state (see Fig. 2). Sequences of such patterns are stored in the AM representing episodes. Therefore, at least in humans and biological agents, SA is based on a **contextual** representation, which is essential for the emergence of the expected SA capabilities as listed in Regazzoni (2020) Table I.

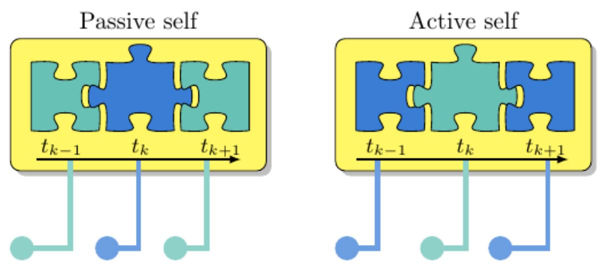

Fig. 2: Two elementary information units depicted in the yellow box correspond to the passive (left) and active (right) self [29]. The passive self unit stores triplets formed by data alternatively acquired from proprioceptive and exteroceptive sensors at different time instants. Proprioceptive data are acquired at time instant t_{k-1} and are followed by data from the environment at time t_k captured by exteroceptive sensors. They cause a change of the internal state of the system at time t_{k+1} that is monitored by proprioceptive sensors. Vice-versa, the active self elementary unit models the cause-effect relation between the data acquired by exteroceptive and proprioceptive sensors.

Figure 3: regazzoni-2020-multi-sensorial-generative-and-descriptive-self-awareness-models-for-autonomous-systems-fig-2.png

Key element in SA knowledge A dynamic description of agent and environment changes based on their reciprocal states is a key element for the representation of SA knowledge. This is different from many traditional AI systems, where exteroceptive sensory data se- quences are often represented at a primary level *without explicit contextual* information.

Traditional AI systems do not consider contextual data: This is different from many traditional AI systems, where exteroceptive sensory data sequences are often represented at a primary level *without explicit contextual information.* As a **consequence**, high-level processing techniques, for example, classifiers based on supervised labeled learning [30], [31], [32], use implicit contextual information to cluster such data into homogeneous groups.

A problem: Despite the impressive classification performance that can be achieved when testing data and training experiences belong to the same class, the observing artificial agent cannot reliably connect such classification results to its internal dynamical state when performing similar actions to the ones performed during training, simply because its state was not observed and memorized together with the observed exteroceptive data. It is therefore not trivial to use such classifiers as building blocks for an artificial agent due to the limited adaptability.

Damasio [28] proposed dispositional units, i.e., information units representing contextualized state changes of the agent or the environment, for modeling "**cognitive cycles**", i.e., episodes that can be found as the *basis of human self-awareness.* Moreover, he suggests that dispositional units can be hierarchically organized at different levels in the brain, for example for describing *temporal-causal* representations in the activation and processing results of neuron maps dedicated to different goals. Consequentially, an AMs should be hierarchical structured for providing SA models, and thus dispositional units' representations should be defined such that they can be organized in multi-level hierarchies (see Fig. 3).

Autobiographical self Neuroscience observations show that the parts of the human brain storing AMs are linked and can exchange neural signals with other parts of the brain known to be activated within conscious inference processes [33]. The role of such neural maps is to analyze—at different hierarchical levels—proprioceptive and exteroceptive sensorial data originating from the current agent's experiences. The process of *recalling* and *comparing multi-level AMs with respect to current experiences* is an *integral capability* of self-awareness related to inference and anomaly detection, which is defined by Damasio as **autobiographical self** (AS).

Application of AS The AS allows an agent to evaluate whether the current experience matches any episode stored in the AM. Moreover, the AS must provide inference processes to interface with other parts of the agent's brain

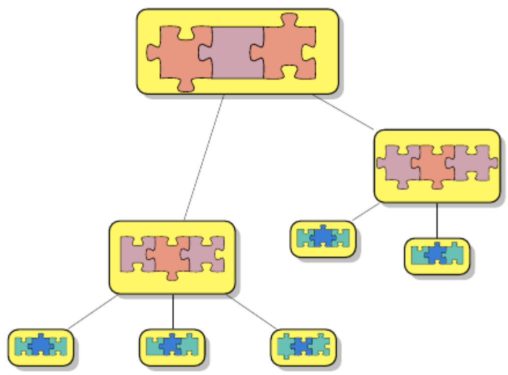

Fig. 3: Hierarchical organization of dispositional units in the autobiographical memory. According to Damasio [28], these elementary information units can be found at different levels in the brain and constitute temporal-causal representations of cognitive processes. This hierarchy expresses experiences at different time scales: directly connecting exteroceptive and proprioceptive data at the leaves and more complex and structured information corresponding to long terms goals at higher-level nodes.

Figure 4: regazzoni-2020-multi-sensorial-generative-and-descriptive-self-awareness-models-for-autonomous-systems-fig-3

(e.g., blocks dedicated to agent's resource planning and control of actuators) to maintain a dynamic stability condition, i.e., homeostasis Regazzoni (2020)[34].

In a SA model, the inference capability implies that activated AMs' dispositional units and currently experienced data elaborated by early neural maps can be managed by the AS inference process to perform, for example, predictions on the agent's future states. Based on the temporal-causal organization of the available episodes stored in the AM, the AS is able to predict future states at multiple abstraction levels by using generative models that represent possible alternative realizations of episodes already experienced adapted to currently observed data. As multiple episodes are stored in the AM, the AS inference processes need to identify models that better match the current experience, which requires the dispositional units' representation in a SA model to inherently provide a discriminative property to assess current data characteristics.

in AI An artificial AS is also required to determine the difference between episodes contained in its own AMs and the current experiences based on an appropriate metric, which can be interpreted as the basis for the abnormality detection capability (see Fig. 4).

In order to assess the matching degree between predictions derived from the dispositional units of the set of potentially applicable episodes and the current observations, the SA agent must apply a computable metric invariant to the sensor modality. In this case, the agent should be aware that an abnormality, i.e., a non-stationary condition never experienced before, is currently present. Damasio does not address which specific computational neural characteristic included in the neural implementation is able to realize such computational behavior. He only suggests that such matching and prediction inference capabilities can be performed by the AS at different abstraction and temporal levels and so enabling an efficient selection of the hierarchical and dispositional representation of AMs episodes).

Rise of emotions It is worth mentioning that for natural agents, Damasio suggests that the integrated SA system composed by AM and AS can also be used as a possible explanation of higher-level human regulatory psychological phenomena such as emotions and feelings [35]. Emotions and feelings can be considered as emergent results of evaluating current experiences based on multi-level hierarchical AMs by means of the AS [25]. For example, fear can emerge from the capability of detecting abnormalities, recognizing that the current experience does not match with past AMs, or it matches with AMs that describe dangerous episodes. Damasio's model implicitly implies that AS outcomes enable an agent to incrementally update in- ternal AM models by coding abnormal experiences into new models as well as to define a SA system that derives inferences invariant to the involved sensor modalities.

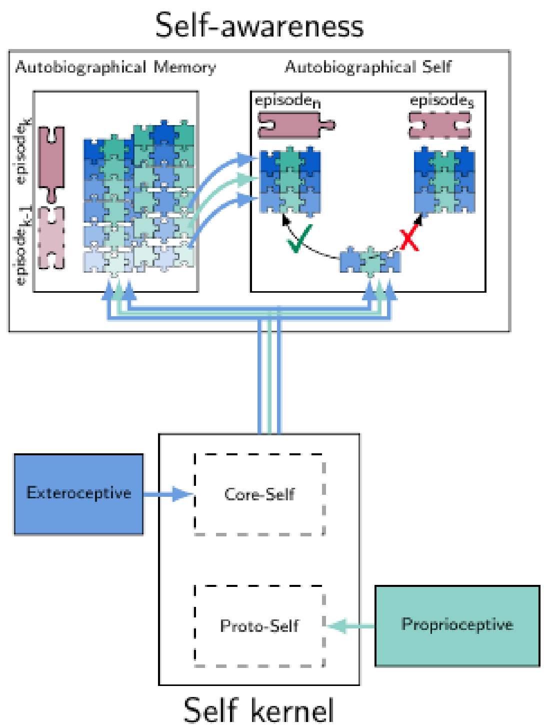

Fig. 4: Autobiographical memory (AM) and autobiographical self (AS) as core components of a self-aware agent founded on Damasio's model. The core-self and the proto-self process exteroceptive and proprioceptive information and store them as dispositional units in the AM. The AS is able to perform inference and anomaly detection based on the stored episodes.

Figure 5: Fig. 4: Autobiographical memory (AM) and autobiographical self (AS) as core components of a self-aware agent founded on Damasio's model. The core-self and the proto-self process exteroceptive and proprioceptive information and store them as dispositional units in the AM. The AS is able to perform inference and anomaly detection based on the stored episodes.

3.2 Haykin's model

In comparison with Damasio's work, Haykin proposes a computational framework of neuroscience observations from an engineering perspective referred to as **Cognitive Dynamic Systems (CDS)** [36]. The proposed CDS model is based on the interactions that a *Cognitive Controller* (CC) part of a CDS has to maintain at multiple levels of abstractions with a *Cognitive Perceptor(CP).*

The CP processes exteroceptive information coming from the environment at different hierarchical levels and can be seen as a hierarchical probabilistic filter, generating environment descriptions at different abstraction levels.
Beyond providing information to higher levels, such a filter generates hierarchical feedback information to the CC, which in turn computes commands to actuators that are characterized by uncertainty. The CC block of Haykin's model is described as a top-down structure generating outputs towards lower levels. *At the bottom layer*, it directly generates outputs for the *actuators.*
The *Probabilistic Reasoning Machine* (PRM), introduced in a joint paper with J. Fuster [37], organizes probabilistic information coming from the CP, i.e., percepts and errors (prediction and update processes), together with information coming from the CC, i.e., planned actions with its related uncertainty. Such an organization is performed over time. As can be seen in Fig. 5, Haykin's model does *not directly employ proprioceptive* sensory data but uses internal strategies to generate commands towards actuators. The **main goal** of the PRM is to maintain a meta-level representation of the *perception-action* cycle based on switching and adapting the behavior of CP and CC.

The relation between Haykin and Damasio As the control strategies embedded in the CP are hypothesized to maintain homeostasis, i.e., a dynamic equilibrium between the agent's state and the changes in the environment, the PRM is contributing to the continuous regulation of agents' processes by providing switching suggestions to CP and CC. Those suggestions are implicitly based on the knowledge that an agent must have been learned from experiences, and it is represented within the PRM. **In this sense, the PRM block is strictly related to Damasio's SA model**, as it has to process representations of actions and percepts organized in a **temporal-causal** order, and the *PRM block can be related to AM and AS as SA* model capabilities.

The PRM elementary representation requires organizing perception and actions into data structures capturing causal and temporal interactions between the agent's actions and percepts originated from the environment. Dispositional information units, as described by Damasio, also represent such interactions but are different because the state of the agent is directly observed by itself. This can be considered either as feedback for the SA agent to evaluate the outcome of commands it has sent to actuators and lower-level blocks or as a multi-sensor source of proprioceptive signals representing the agent state to itself. This concept is exploited in Regazzoni (2020) SA model (depicted in Fig. 10). If this second view is taken, a modified PRM can be considered as a structure where

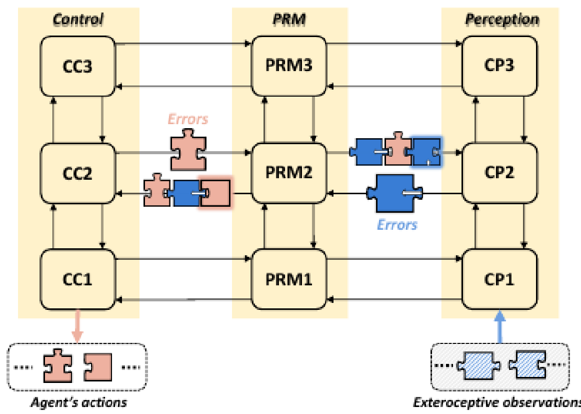

Fig. 5: Hierarchically structured probabilistic reasoning machine (PRM) adopted from [37]. The cognitive perceptor (CP) processes exteroceptive information ("percepts") at different hierarchical levels, and the cognitive controller (CC) generates outputs towards the actuators in a top-down structure. The PRM organizes probabilistic information from perception and control by data structures capturing causal-temporal interactions that are similar to "dispositional units".

Figure 6: Fig. 5: Hierarchically structured probabilistic reasoning machine (PRM) adopted from [37]. The cognitive perceptor (CP) processes exteroceptive information ("percepts") at different hierarchical levels, and the cognitive controller (CC) generates outputs towards the actuators in a top-down structure. The PRM organizes probabilistic information from perception and control by data structures capturing causal-temporal interactions that are similar to "dispositional units".

appropriately modified dispositional units are processed based on hierarchical filters that work on proprioceptive feedback and in a bottom-up way also in the CC.

In Haykin's approach, the *focus is on control* instead of SA. Therefore, the PRM is designed to make a CDS capable of using interactive behavioral rules to switch among different perceptions and action modalities adaptively. Such inferences can drive actions that the agent's sensors and actuators should accomplish anticipatorily by activating available models in the PRM memory when it performs a given experience. This allows the CDS of an attention capability towards preferred control or sensing actions in a given homeostatic cycle. In the case of the SA model discussed here, the agent has no direct knowledge of the control strategy actions that are generated by its own decision-making subsystem, but it can observe and fuse their outcomes through parallel proprioceptive feedback from sensors coupled with exteroceptive environmental observations. Haykin's model, however, suggests how a SA model can provide useful data to adapt decision making processes behaviors at homogeneous hierarchical levels to the PRM. For example, SA can share with decision-making estimates of present and predicted contextualized states as well as errors and deviations of models describing previous agent experiences.

Furthermore, Haykin's model facilitates identifying temporal variations of uncertainty associated with actions and percepts, which is a key aspect for addressing a proper computational framework for the CDS design, i.e., a PRM in Regazzoni (2020) SA model. Moreover, the organization of a PRM at multiple abstraction levels is coherent with the hierarchical characteristics of Damasio's dispositional units. In this case, actions and percepts as temporal aggregations (equivalent to dispositional units) at different hierarchical levels can describe the joint state of lower-level parts of the agent body down to the directly observed proprioceptive and exteroreceptive characteristics of the agent and the environment. Although Haykin's approach does not provide a specific probabilistic model for uncertainties and dependencies, it proposes a Bayesian framework to model uncertainty and causality and make inferences computationally, e.g., parametric conditional probability models.

Although the goal of Haykin is not to specify a univocal PRM model but to provide a generic framework for CDSs, his model is essential for addressing the main techniques for SA in artificial agents. His work then suggests that SA models origi- nating from a computational domain should be associated with an appropriate calculus of uncertainty propagation. In [38], a CDS inspired by such a probabilistic approach uses a simple PRM unit that allows a vision system on a mobile platform to make inferences about future states. Moreover, Haykin's work does not directly provide a unitary view of the techniques that could be used to store a coherent multi-level generative and discriminative PRM's knowledge. Nonetheless, Regazzoni (2020) work integrates Haykin's viewpoint on uncertainty and makes a relation between the perception-control blocks and Damasio's theory, which includes AM and AS and dispositional units. Regazzoni (2020)Fig. 10 displays a revisited block scheme of Haykin's model.

3.3 Friston's model

Another relevant cognitive framework for SA that aims at establishing links between neuroscientific observations and computational models is the one proposed by Friston [39], [27]. Here, Bayesian dynamical systems are the computational tool that facilitates an uncertain and hierarchical self- coherent representation to describe and generate simulations of inferences performed by the human brain utilizing neuron firings. Friston's approach is innovative in the context of developing self-aware models for artificial agents due to the following characteristics: i) It formally relates a statistical me- chanics' optimization framework, that can be summarized as free energy and variational based reasoning, with Bayesian inference. It founds a theoretical domain for describing SA knowledge and models (in the AM) as well as inference (in the AS). ii) It proposes the concept of generalized states (GS) to develop a class of computational and hierarchical Bayesian filters that we use to embed representation and inference over dispositional temporal knowledge.

The good regulator theorem [35] states that "every good regulator of a system must be a model of that system". In this sense, SA models can be considered as joint discriminative and generative models that contribute to the regulation of an artificial agent representing an adaptive code of the system itself and its incremental experiences at the same time. The free energy principle represents an optimization criterion that can be related with a variational computational framework to both define and discover optimal SA models from a given set of dynamic experiences, i.e., available data sequences originating from exteroceptive and proprioceptive sensors.

In [27], Friston suggests that establishing an equivalence between a probabilistic and a mechanical statistics' representation of the dynamic equilibrium in the sensed internal state and the contextual environment allows one to explain the observed neuron firing in the human brain through the free energy concept. He further shows that Bayesian inference is an equivalent way to do so. As SA in humans is based on brain inference processes, probabilistic dynamic representation and inference models are good candidates to form the **language** for expressing a SA model in an artificial agent as well. Such models must be capable of including temporally ordered descriptions of contextual dependencies between proprioceptive and exteroceptive variables.

The variational computational representation and inference techniques he proposes clarify how different models can describe statistically different sensorial experiences that can be seen as trajectories in generalized spaces. The models can also provide explicit measurements to evaluate or to discriminate best-fitting models of new observed sequences. At the same time, the same models, that can include multiple conditionally connected random variables at different hierarchical and temporal levels, are capable of predicting multi-level, temporal data series characterized by the same statistical properties of the training experiences from which they can be learned (thanks to their generative nature). Such a model, if used within the SA model of an artificial agent, together with appropriate learning techniques, can facilitate incremental model creation. An AS can so actively memorize incrementally generative and discriminative new models in the AM by processing sensorial experiences. Such models have to capture different causality interactions between the agent and the environment and serve as the basis for *symbolic* descriptors in an artificial SA agent.

The SA capability of abnormality detection can also be explained with Friston's model, i.e., the free energy that models generate when the AS compares them to a new experience. A metric can be defined to evaluate the amount of abnormality which is related to a particular component of free energy, describing the orthogonal perturbations to the dynamic equilibrium condition described by the model. Such a metric enables the AS to rank the abnormalities measured from a set of AM models, so relating the discriminative SA property of the model to the abnormality detection and inference capabilities.

Contribution ii) to the SA model definition is a specific class of Bayesian filters, namely GS filters. Friston et al. [40] explain that thanks to such filters, active and variational Bayesian inference techniques can be obtained with better performances. GSs describe a class of trajectories in terms of generalized coordinates of motion. The resulting model can be shown to better describe the dynamical nature of the pattern in terms of temporal and causal explainability of dependencies among states as well as computational benefits. In a SA model, hierarchical Bayesian models derived from GS filters can provide the description language for the AM, depicted as individual "pages" in Fig. 6. Regazzoni (2020) shows how these models can be learned by observing proprioceptive and exteroceptive data series both independently and in a coupled way that is oriented to represent their dispositional nature. Such filters can be used both as generative and discriminative models, and they can be good candidates to be related to the "good regulator" code [35], as they can represent the rules that describe the agent, as well as such rules, can be used by the agent to predict the dynamic contextualized behavior where it is acting.

Coupled proprioceptive and exteroceptive signals of GS filters can efficiently represent Damasio's dispositional units in a SA model. For example, a Switching Dynamic Bayesian Network (DBN) [41], [42] uses multi-level discrete and con- tinuous generalized states as variables that will be further discussed in the following sections.

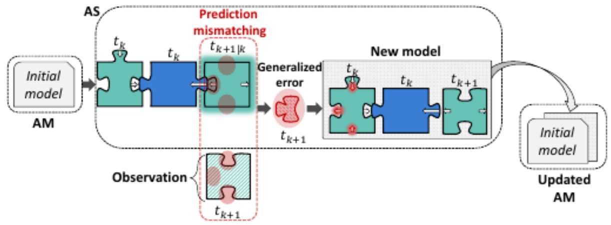

Fig. 6: An **initial** model is employed to predict passive self states, i.e., glowing green puzzle pieces at instant $t_{k+1|k}$. Errors from such a model are utilized to **create** new models, depicted as new *pages* in the **AM** structure. Such new models minimize the free energy between the agent's **inferences** and the observed data. Note that the same logic can be applied to an active self (blue-green-blue puzzle pieces).

Figure 7: Regazzoni (2020) Fig. 6

Dynamic Expectation Maximization (DEM) filters [40] are hierarchical parametric GS filters that are here used to derive coupled GS-DBNs. These filters have been shown to jointly perform parameter, hyperparameters, and GS estimations within a continuous variable Bayesian network that is by itself a fully continuous DBN. However, discrete variables are needed in a SA model too. Such variables are used to represent different models (i.e., different pages in the AM structure) and to provide finer level discriminative descriptions of learned models to determine a different class of probabilistic dependencies within an episode, useful both for generative and discriminative purposes. Coupled GS-DBNs are, therefore, better-suited filters than DEM, in particular when all model properties are necessary to reach the SA capabilities (see Fig. 7).

4 SIGNAL REPRESENTATION AND MODEL LEARNING

4.1 Bio-inspired SA Model

The main difference between Haykin's and Damasio's mod- els is that the objective of the latter lies in the bottom-up explanation of neuroscientific observations, while the first one focuses on the definition of control within a CDS. Similar to Haykin's model, Friston's approach is based on a Bayesian and computationally efficient approach for SA representations. Friston's approach is more focused on a bottom-up joint analy- sis of proprioceptive and exteroceptive signals, while Haykin's model aims at providing a framework for defining compu- tational aspects of perception-action cycle decision making outcomes towards actuators in a CDS. A computational SA model for an artificial agent

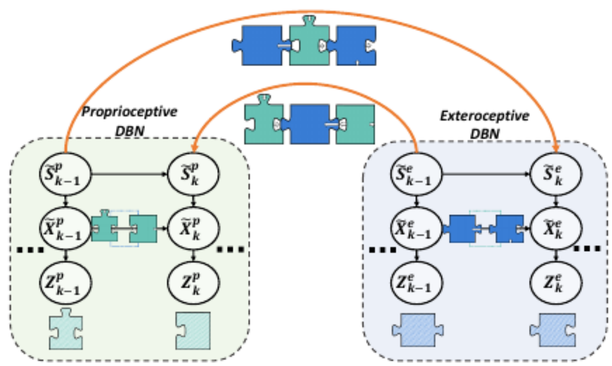

Fig. 7: Proprioceptive DBN (P-DBN) represented in a green block and exteroceptive DBN (E-DBN) represented in a blue block are connected by orange links that encode the agent's **contextual information**. Each DBN (P-DBN and E-DBN) performs continuous $\tilde{\boldsymbol{X}}$ and discrete $\tilde{\boldsymbol{S}}$ inferences. This coupling facilitates to model **interactions** between multi-sensory data and perform inferences within a contextual SA framework. Observations are represented as Z, and k encodes time instances. Proprioceptive and exteroceptive information is indexed as p and e, respectively.

Figure 8: Regazzoni (2020)Fig. 7

TABLE II: Comparison of inherent self-awareness properties of the presented bio-inspired self-awareness theories.

Self-awareness properties	**Damasio's model**	**Haykin's model**	**Friston's model**
Generative modeling	Dispositional units are facilitated to make non-probabilistic predictions of future agent's states based on a top-down approach	Predictions of next actions and exteroceptive states performed by the PRM	Predictive GS probabilistic models
Discriminative modeling	Not considered	Not considered	Focused more on filtering than on semantic labeling of experiences
Interactive	It includes dispositional units but interactions are not explicitly explained	PRM relates information between exteroceptive data and agent's actions	Self-organization in agents explained as system of GS filters related to agent actions and sensory perceived environment
Hierarchical modeling	It considers several abstraction levels ranging from raw observations to feelings/emotions	Multilevel representation of control, PRM and perception, see Fig. 5	Continuous variables in upper inference levels parameterize predictive models in lower ones
Temporal reasoning	Dispositional units relate present states with future ones	Temporal dependencies between control and environment perception during time	In Bayesian DEM filters different temporal reasoning at abstraction levels of parameters and GSs
Uncertain reasoning	Not considered	Bayesian reasoning	Equivalence of active Bayesian inference and attractors in statistical mechanics

Figure 9: Regazzoni (2020) Table II

can be obtained by merging different aspects of the three frameworks. Such SA model should include SA properties as discussed in Section I and enlisted in Table II.

5 Levels of self-awareness

Lewis et al. (2011) II-A

- Ecological self
- Interpersonal self
- Extended self
- Private self
- Conceptual self

6 Private/Public self-awareness - Subjective/Objective

Lewis et al. (2011) II-B

References

Raja Chatila, Erwan Renaudo, Mihai Andries, Ricardo Chavez-Garcia, Pierre Luce-Vayrac, Raphaël Gottstein, Rachid Alami, Aurélie Clodic, Devin Sandra, Benoît Girard, and Mehdi Khamassi. Toward self-aware robots. 08 2018.

Peter R. Lewis, Arjun Chandra, Shaun Parsons, Edward Robinson, Kyrre Glette, Rami Bahsoon, Jim Tørresen, and Xin Yao. A survey of self-awareness and its application in computing systems. In *Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems, SASOW 2011, Ann Arbor, MI, USA, October 3-7, 2011, Workshops Proceedings*, pages 102–107. IEEE Computer Society, 2011.

Carlo Regazzoni. 2020.