

Toward artificial collective self-awareness

Mohammad Rahmani

Decide Doctrol Group

July 1, 2020

Self-awareness (SA) - Definition

- The capacity to become the object of ones **own attention**, which arises when an agent focuses **not only** on the **external environment** but also on the **internal milieu** [9].

SA - Definition (Literal)

- Giving the ability to computers to program themselves in unseen circumstances.
- Casual-temporal inference of agents behavior on the environment and vice-versa.
 - If I do x , what will happen outside? (Part of behavioral Active awareness)
 - If y happens outside, how would it affect me? (Part of behavioral Passive awareness)

SA - Levels

See [7] - Ordered from basic to advanced

- **Ecological self** The most basic, referring the ability of an agent to react to a stimuli.
- **Interpersonal self** Awareness of external interaction such that limited adaptation to performance of basic homeostatic tasks is achieved.
- **Extended self** Permit reflection of interactions over time. The organism is aware of the existence of past and future.
- **Private self** The agent can process more advanced information concerning itself,
- **Conceptual self**: The capability of constructing and reasoning about an abstract symbolic representation of itself (AI's goal).

SA - Aspects

See [7]

- **Identity Awareness:** The ability to recognize and model the identity of agents.
- **State Awareness:** The ability to model and recognize the states of oneself, the world or other entities within it
- **Time Awareness:** The knowledge of past or potential future basic stimuli
- **Interaction Awareness:** The ability to taking into account casual patterns of interactions between entities.

SA - Aspects (2)

See [7]

- **Behavior Awareness:** The ability to model the internal behavior of the system or behavior of external entities.
- **Goal Awareness:** the ability to conceptualize the internal factors that drive the behavior, such as a systems goals, objectives, and constraints.
- **Belief Awareness:** Things believed to be true by a system which do **NOT** need to capture the notion of time.
- **Expectation Awareness:** Combines belief awareness and time awareness, to form models that express what the system or others believe about how the world will unfold over time

Why SA in AI?

The driving motivation for the transfer of biological SA concepts to artificial systems is to improve:

- autonomy
- robustness
- scalability

In intelligent Agents (IA).

SA IA Essential requirements

See [10]

- **Initialization:** Initial knowledge from which an agent starts building its own memories (Training phase. *Techniques:* Random walk, Driving an autonomous vehicle a few times human agent. etc). **Biology:** Implemented into genes.
- **Memorization:** Capability to store and retain information **Biology:** Brain's capability to fire group memory related neurons in case of facing familiar patterns of stimuli.
- **Inference:** Ability to predict own future states **Biology:** See "Predictive coding theory" in [11]

SA IA Essential requirements 2

- **Anomaly detection:** Capability to recognize observations which don't match episodes of memory. **Biology:** Brains ability to send it's predictions to low-level sensory regions to test them against existing memories.
- **Model creation:** Capability of generating models that encode previous experiences for future predictions. **Biology:** In brain, internal models get adjusted so that the predicting error gets suppressed [3].

SA IA Essential requirements 2

- **Decision-making influence:** The ability to generate signals that can be employed by the agents control system such that its actions are self-monitored dynamically. **Biology:** Muscles move based on commands from the brain [4]. Nerve cells in the spinal cord, called motor neurons, enable to convey and evaluate the brains commands to the muscles.

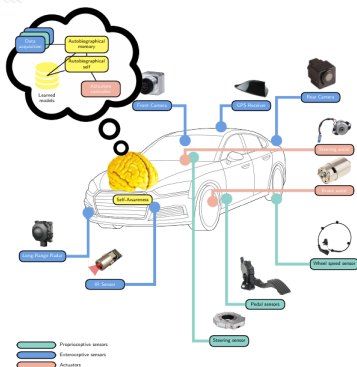
SA - Existing biological models

- Damasio [1]
- Haykin [5]
- Friston [3]

Damasio model, more or less presents an architecture for self-awareness but doesn't present a computational approach. The two others, have to some extent adopted Damasio's model but have also presented mathematical models to justify how observation of an external stimuli leads to decision making.

Damasio SA

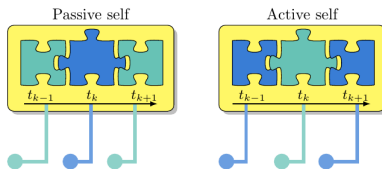
- Lets divide observations of an agent to **exteroceptive** and **proprioceptive** observations.



See [10]

Damasio - Dispositional units

- Lets make the contextually put together proprioceptive and exteroceptive data over the course of time in two ways:
 - Passive: An exteroceptive piece of data is surrounded by two proprioceptive pieces of data
 - Active: A proprioceptive piece of data is surrounded by two exteroceptive pieces of data

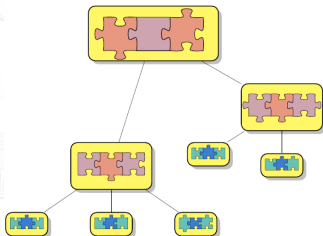


See [10]

Damasio - HDU

Hierarchical Dispositional Units (HDU)

- To extract casual-temporal inferences, similar to brain, dispositionl units are contextually placed in one another in a hierarchy.



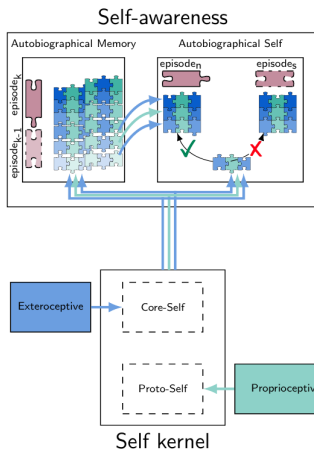
See [10]

Damasio - AM

Autobiographical Memory (AM)

- Each set of dispositional units which contribute to building an experience is stored as an episode
- The collection of all episode forms the architecture of memory in the form a book which each page presents an episode.

Damasio - AM



See [10]

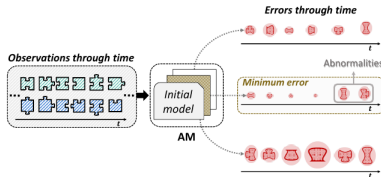
Damasio - Abnormality

- Lets open the system's memory to observation through the time.
- Try to find a match for the set of observations in all AM pages with some acceptable noise tolerance (Most models use Hellinger metric to measure the distance between model's prediction and the set of observations)
 - If there is a prediction(model/episode in AM) that falls below the tolerance for that set of observations then activate actuators as the models says.
 - If none of the predictions doesn't fall below the tolerance then its time to add a new episode (i.e page) to AM.

So another definition of self awareness is:

- Learning from large abnormalities

Damasio - Abnormality



See [10]

Computational SA approaches

- Bayesian Model (BM): To model causality
- Dynamic BM (DBM): To add temporality to causality model
- Coupled DBM (To model dependence between Extroceptive and Proprioceptive events)
- Generalized filtering (A non-linear space state Bayesian filtering) (To predict future state of an agent)

Collective SA (CA)

Examples:

- Bee colony
- Ant colony
- human immune system

In all above models no individual agent shares all the information with other agents but through development of a local interaction self-awareness which is distributed among all. Ants use pheromone samples left by other ants to locate the food. **Hence collective self-awareness is always decentralized and distributed.** [8]

Computational CA application

Examples:

- Agent collision avoidance
- Traffic jam avoidance

Benefits:






- Improves scalability
- Reduces computational complexity through local-symbolic interaction
- Addresses heterogeneity by creating semantic fields

Computational CA approaches

See [10, 6, 2]

- Lets correlate sets of observations of each agent with its true status
- Lets cluster the space of those states such that addition and removal of classes is possible during the time the procedure is taking place (Using Growing Gas Neural Networks (GGN)) - K-means and Self-organizing map doesn't help in this case.
- Use the attained clusters as letters of words to make the state space discreet.
- Propagate the discreet states between the agents instead of real-data.
- Hellinger distance for abnormality detection
- Generate new models from large abnormalities.

References

-  Antonio Damasio. *The feeling of what happens : body and emotion in the making of consciousness*. New York: Harcourt Brace, 1999. ISBN: 0156010755.
-  Lucio Marcenaro David Martin Arturo de la Escalera Carlo Regazzoni Divya Kanapram Pablo Marin-Plaza. "Cognitive Dynamic Systems: Perception-action Cycle, Radar and Radio". In: 2019.
-  Karl J. Friston. "The free-energy principle: a unified brain theory?" In: *Nature Reviews Neuroscience* 11 (2010), pp. 127–138.
-  V. Gallese G. Rizzolatti L. Fadiga and L. Fogassi. "Premotor cortex and the recognition of motor actions". In: *Cognitive Brain Research*. Vol. 3. 1996, pp. 131–141.
-  Simon Haykin. "Cognitive Dynamic Systems: Perception-action Cycle, Radar and Radio". In: 2012.