Towards a Brain Inspired Model of Self-Awareness for Sociable Agents

Budhitama Subagdja
Joint NTU-UBC Research Centre of Excellence
in Active Living for the Elderly (LILY)
Nanyang Technological University, Singapore
e-mail: budhitama@ntu.edu.sg

Ah-Hwee Tan
School of Computer Science and Engineering
Nanyang Technological University, Singapore
e-mail: asahtan@ntu.edu.sg

Abstract

Self-awareness is a crucial feature for a sociable agent or robot to better interact with humans. In a futuristic scenario, a conversational agent may occasionally be asked for its own opinion or suggestion based on its own thought, feelings, or experiences as if it is an individual with identity, personality, and social life. In moving towards that direction, in this paper, a brain inspired model of self-awareness is presented that allows an agent to learn to attend to different aspects of self as an individual with identity, physical embodiment, mental states, experiences, and reflections on how others may think about oneself. The model is built and realized on a NAO humanoid robotic platform to investigate the role of this capacity of self-awareness on the robot's learning and interactivity.

Introduction

Intelligent systems have become popular and applied as daily companions that interact closely with humans. Beyond user-initiated modes of interactions like menu selection or direct manipulation, emerging applications and systems are becoming more human-like to interact naturally with the user. Some of them have the capacity to converse in natural language and learn the user preferences (e.g. virtual assistants like Siri, Cortana, or Now!). However, they do not capture what they have been doing, their experiences so far from the beginning of their uses, nor their relationships with the users. Thus, they do not really reflect on what they learn about the users in order to improve their own interactivity.

In an imaginary futuristic scenario, a robotic companion can be viewed as a unique individual that has its own identity, personality, and social life beyond simply an application or a toy that performs routine tasks. People may ask for the robot's opinion based on its actual reflection on its own thought, feelings, and experiences beyond answering specific questions or advising prescribed recommendations. The objective of this paper is a start to move towards that direction in realizing self-awareness for a computational agent as in the above scenario.

Known in psychology as the capacity to put oneself as the focus of attention (Duval and Wicklund 1972), self-awareness is crucial for an intelligent agent to handle the natural interaction with people. Existing computational models of self-awareness mostly focus on modeling specific self-embodiment at the physical aspect (Bongard, Zykov, and Lipson 2006; Hart and Scasselati 2012; Stoytchev 2011) to recognize and map one's self in the mirror (Hart and Scasselati 2012; Stoytchev 2011) or to robustly control one's movement regardless physical damages or impairments (Bongard, Zykov, and Lipson 2006). A higher level model of self-awareness has also been proposed that covers the social aspect of the robot, but it is made only to deal with a specific task of epistemic logical reasoning (Bringsjord et al. 2015). Despite the above progresses, making a comprehensive but practical model of self-awareness remains a great challenge. Few, if any, have realized self-awareness in a robot or an agent to interact fluently with people and learn from one's own experiences in order to understand one's self in the social context.

In this paper, we propose a model of self-awareness that covers five aspects of individual including identity, physical, mental, experiential, and social. With a holistic representation, the awareness is not just limited to the first-person impressions but also from a third-person perspective. This paper also presents a hybrid neural architecture of self-awareness that can dynamically hold multiple possible selves together in a hierarchical manner. This model facilitates not just the integration of different components in the system but also their combination and arrangement to create possible selves.

The model is embedded into a NAO robotic platform that can roam the environment while encountering and making conversations with people. It has some basic prior self-knowledge and concepts at the start that can be refined and added incrementally through dialog and observation. The capacity of self-awareness is demonstrated to incorporate dialogs with knowledge and memory of past experiences.

Related Work

Self-awareness is known in social psychology as the capacity to pay attention to oneself including the attention to what one is thinking, doing, and experiencing (Duval and Wicklund 1972). In Neisser (1997), self-awareness can be characterized in different levels, including *ecological*, *interpersonal*, and *conceptual* levels.

Self-awareness for computational agents has been realized mostly in robotics or virtual agents as the capacity to automatically model one's own bodily structure and self-identification. Learning the physical properties of bodily actions through "motor babbling" has been applied to a robot that detects whether the image is a mirror reflection of itself (Stoytchev 2011). A more advanced model of self-awareness enables the robot to make use of the acquired self-model to project its current bodily actions to its mirror reflection (Hart and Scasselati 2012). The acquired model of self can also be used to control the robot movement and to continuously adapt the model, making it robust despite physical damage and impairment (Bongard, Zykov, and Lipson 2006).

Besides modeling one's embodiment, a feature of self-awareness has also been applied to demonstrate an epistemic logical reasoning about knowledge of one's self and the others (Bringsjord et al. 2015). The robot demonstrates the ability to infer the state of self after observing the states of other surrounding robots. Although this later work demonstrates self-awareness in a higher-level reasoning rather than low-level bodily knowledge, it is still made to deal with very specific logical tasks.

Conceptually, self-awareness has been considered to be a part of a cognitive agent architecture (Sloman and Chrisley 2003; Sun 2007; Samsonovich and Nadel 2005) which sometimes is considered to be a meta-cognition that manages introspection. Others suggest that the self should be represented explicitly and exchanged among memory modules. In a survey about self-awareness in computing system (Lewis et al. 2011), it is suggested that self-awareness can be ascribed to a computing node if it maintains the information about its own internal states (private self-awareness) and has sufficient knowledge to determine how it is perceived by other parts of the system (public self-awareness). Similar to the last conceptualization, our model adopts the view that self-awareness can be achieved if all aspects including the private and public awareness are considered.

Self-Awareness

Self-awareness can be defined as a capacity to introspect. One may be aware of something or someone in the world, of an imaginary object one has seen previously, or even of an abstract thought (see Figure 1(i)). However, introspection must be the awareness about oneself as an individual ascribed with personification. Self-awareness is produced based on a constructive process. Evidence from neuropsychology suggested that typically awareness of self in physical or bodily level is mentally made up (Ramachandran 2011).

As a constructive process of mentally representing oneself, self-awareness occurs occasionally and can be changing dynamically depending on the situation, context, or the one's personality. Since an introspection typically holds from a first-person view, one can represent oneself as the subject in the introspection who is doing something in the world. This first-person introspection can be called *subjective self-awareness* (see Figure 1(ii)). When interpersonal aspect is considered, oneself can be positioned as the object being observed from a third-person perspective. The later can also be called *objective self-awareness*(see Figure 1(iii)).

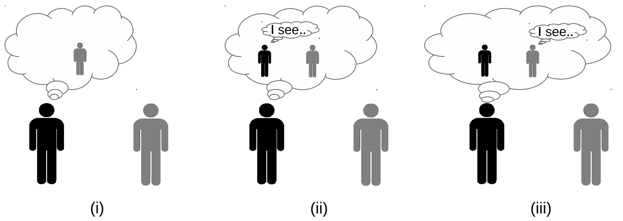


Figure 1: (i) No self-awareness occurs if one's self is not represented even though one may pay attention to someone else; (ii) In *subjective self-awareness*, one is represented as the subject from the first-person; (iii) In *objective self-awareness*, one is represented as observed from someone else or a third-person.

Aspects of Personified Individual

In this paper, we define five aspects of personified individual that can be ascribed to an entity so that it can have the self awareness capacity.The aspects are as follows:

- **Identity**. Identity is an attribute ascribed to one individual so that one can be identified, characterized, and distinguished from another individual. Included in this aspect are the idiosyncratic conventional identification (e.g. name, registry number, owner, guardian, parents, place of origin) and the personal attribute that conceptually characterizes the individual (e.g. intellectual, professional, youth, hot-tempered, generous). Attributes indicating whether the individual is a subject or an object are also included.
- **Physical**. One can access information and knowledge about one's own embodiment, sensations, objects perceived in the environment, and actions, either currently on-going at the present or potentially occurring in the future. This aspect corresponds to the ecological self in the level of self-awareness as prevously mentioned (Neisser 1997).
- **Mental**. Considered as the central part of an individual, this aspect includes the ascription of mental attributes like goal (to have something to achieve), intention (to perform some actions to achieve a goal), thought (to have attention to something imaginary, predicted, hypothesized, or reminisced), and feeling (angry, sad, afraid, disgust, embarrassed, pride). This mental attribution fuses and integrates different attributes in every aspect of self which explains the use of the term reflection. Thus, the attribution can be hierarchical or arranged in complex interrelations among information items.
- **Experiential**. One can be ascribed with information regarding one's own experiences in the past. The remembered information can be used to project oneself to the future as a prediction or hypothesis. The projection can also be imaginary or unreal. Similar to the mental attribution, the projected self constitutes attribution from the other aspects of individual.
- **Social**. In this aspect, one is related to other individuals. In a simplest form, this other individual can be ascribed with

social impression (e.g. loved, hated, best friend, brother, sister, parent). In a more complex form, the individual can be ascribed with mental attributes in a nested structure and projections from one individual to another which can realize *objective self-awareness*. This aspect may also include social rules or norms that commonly drive the behavior of a person.

Formalization

To be more precise, self-awareness can be formulated as tuples structured in nested hierarchies and sequences. The tuples represent what is in an agent's mind. More formally, an awareness χ_a of agent a can be defined as a set

$$\chi_a = \{\varsigma_{a_1}, ..., \varsigma_{a_m}\}, \tag{1}$$

where an element ς_{a_i} is an ordered list of introspection units of an agent a_i that represents one's own introspection as a narrative or a story in a sequential order. Thus, $\varsigma_{a'}$ of agent a' is a sequence or a finite ordered list

$$\varsigma_{a'} = (\psi_{a_1 1}, ..., \psi_{a_n p}), \tag{2}$$

in which an item $\psi_{a_j k}$ is the introspection unit of agent a_j at the kth position in the sequence. An introspection unit ψ_{a^*} of agent a^* can be defined as

$$\psi_{a^*} = (r'_{a^*}, \{r_{a_1 1}, ..., r_{a_q s}\}), \tag{3}$$

where r'_{a^*} denotes a reflection of agent a^* in which a^* is the subject. Other reflections in $\{r_{a_1 1}, ..., r_{a_q s}\}$ are the objects in a^*'s introspection.

A reflection $r_{a^‘}$ of agent a‘ can be defined as a tuple

$$r_{a^‘} = (id_{a^‘}, phys_{a^‘}, mind_{a^‘}, ctxt_{a^‘}, soc_{a^‘}), \tag{4}$$

wherein $id_{a^‘} \in \mathcal{I}d$, $phys_{a^‘} \in \mathcal{P}y$, $ctxt_{a^‘} \in \mathcal{C}x$, and $soc_{a^‘} \in \mathcal{S}c$ are tuples ascribed to agent a‘ based on the personification aspects of *identity*, *physical*, *experiential*, and *social* respectively. $\mathcal{I}d$ is the set of possible identities to characterize an individual agent. $\mathcal{P}y$ consists of possible characteristics of physical (bodily) states, perceived objects, and environmental conditions. $\mathcal{C}x$ is the set of possible description of context (time, location, situation). $\mathcal{S}c$ consists of possible social attributions and relationships.

On the other hand, $mind_{a^\star}$ of an agent $a^\star$ can be defined as a tuple that

$$mind_{a^\star} = (m_d, \chi^l_{a^\star}), \tag{5}$$

where $m_d \in \mathcal{M}d$, and $\mathcal{M}d$ is the set of possible mental attributes to ascribe (e.g *belief*, *achieve*, *intend*, *imagine*, *recall*, *predict*, *anger*). $\chi^l_{a^\star}$ is agent $a^\star$'s awareness as defined in (1), but at l level of mental hierarchy so that it is structurally nested. l is the level of depth of awareness such that $\chi^1_{a^\star}$ is the first level of one's ($a^\star$'s) awareness about an individual. In this paper, the expression $\chi^1_{a^\star}$ is simplified as $\chi_{a^\star}$.

The awareness model above represents social awareness in general, not specifically self-awareness. An agent a may have an introspection unit that is not referring to a's self at all. In that case, self-awareness can be defined based on which individual within the χ representation should be referred to as a subject or object.

Definition 1 Subjective self-awareness *is an awareness* χ^l_a *of an agent* a *that, for* $l = 1$*, there is an introspection* $\psi_{a_i n}$ *with subject* r'_{a_i} *that* $a_i = a$ *or* a *is the subject (first-person).*

The above definition ensures that one must be the subject of introspection (the 'I' as the subject of experiences).

Definition 2 Objective self-awareness *is an awareness* χ^l_a *of an agent* a *that, for any level* l *of awareness, there is an introspection* $\psi_{a_i n}$ *wherein reflection* r_a *of agent* a *is a member of its reflection objects (third person).*

The definition above suggests that a's objective self-awareness can be achieved if at least in one of a's reflections, some individual a_i refers a as the object (me) in a_i's mind.

Computational Architecture

At a moment of time, an agent may create and hold an awareness χ in one's mind. With χ, the agent can relate to another individual and organize one's thoughts as a hierarchy of different agents associated with their own minds. Conforming with neuropsychology, it has been suggested that the self-awareness capacity is strongly related to social faculty in the brain and the capacity to organize thought and actions (Pfeifer, Lieberman, and Dapretto 2007).

Based on χ as a dynamic abstract representation, in this section, the model of awareness is presented more specifically as a practical brain-inspired computational model that exhibits self-awareness.

ARTSELF: Neural Model of Self-Awareness

In this paper, we propose ARTSELF as a model of neural network that can exhibit self-awareness. ARTSELF is based on fusion ART (Tan, Carpenter, and Grossberg 2007) which is a variant of ART (Adaptive Resonance Theory) network (Grossberg 2013) that incorporates multiple input (output) neural fields.

Fusion ART

ART applies bi-directional processes of categorization (bottom-up) and prediction (top-down) to find the best matching category (resonance). Learning occurs by updating the weights of connections between the top and the bottom layer of neural fields at the end of a resonance search cycle. ART may also grow dynamically by allocating a new category node if no match can be found.

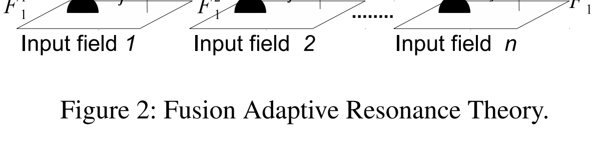


Figure 2: Fusion Adaptive Resonance Theory.

Specifically, as shown in Figure 2, fusion ART has n input fields (F^k_1 is the kth input field) and one category field F_2. Let $\mathbf{x}^k$ and $\mathbf{w}^k_j$ be F^k_1 activity vector and weight connection vector between F^k_1 and jth node in F_2 respectively.

Resonance search is a process of finding a node J in F_2 such that

$$T_j = \sum_{k=1}^{n} \gamma^k \frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{\alpha^k + |\mathbf{w}_j^k|}, \quad (6)$$

$$T_J = \max\{T_j : \forall k, m_j^k = \frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{|\mathbf{x}^k|} \geq \rho^k, \forall j \, \text{in} \, F_2\}, \quad (7)$$

where the fuzzy AND operation $\wedge$ is defined by $(\mathbf{p} \wedge \mathbf{q})_i \equiv \min(p_i, q_i)$, and the norm $|.|$ is defined by $|\mathbf{p}| \equiv \sum_i p_i$ for vector $\mathbf{p}$ and $\mathbf{q}$. $\alpha^k > 0$, $\gamma^k \in [0, 1]$, and $\rho^k \in [0, 1]$ are *choice*, *contribution*, and *vigilance* parameter respectively. Among the parameters, γ^k is the significance of F_1^k and α^k is to avoid division by zero. Vigilance ρ^k, on the other hand, determines how specific the resonance search matches a node with the input.

When a match is found, learning may take place by updating the corresponding connection weights towards the input values. The weights, then, can be read out to any F_1 field to complete some partial input (if any). In this case, the output of the network is at the bottom layer. On other hand, if no resonance is found, that is $\forall j \, \exists k, m_j^k < \rho^k$, an uncommitted node j' in F_2 is allocated to store the input as a new entry which ensures the stability of learnt knowledge. Thus, the network will grow as it encounters novel input patterns.

Fusion ART has been successfully applied to control non-player characters that perform on-line reinforcement learning in a fast-paced real-time first-person shooter video game (Wang et al. 2009).

ARTSELF

The awareness model can be realized with a neural network by making multiple layers of fusion ART wherein the activation patterns on every field in a layer represents the awareness or introspections. In Grossberg (2013), it is suggested that the ART network dynamic can explain how awareness and attention to multiple objects happen in the brain cortical area. In particular, the transient activation of nodes in a category field can be regarded as concepts or objects being attended to (aware of) and held together at a particular moment. The same principle has successfully been applied to demonstrate a transient formation of goal hierarchy for planning (Subagdja and Tan 2012), episodic memory (Wang et al. 2012; Subagdja and Tan 2015), and autobiographical memory (Wang, Tan, and Miao 2016).

Figure 3 shows ARTSELF as a three-layered network of fusion ARTs with one stacked on top of another. The exception is the top layer (F_4) that has a direct connection to one of input fields (F_1^{id}). The bottom-most fusion ART consists of input fields F_1^{id}, F_1^{phys}, F_1^{mind}, F_1^{ctxt}, and F_1^{soc} which corresponds to reflection tuple r_a defined in (4). A reflection r may consist of complete or partial description of individual at a certain physical, mental, social, and/or experiential situation. r can be expressed as input vectors ($\mathbf{x}^{id}$, $\mathbf{x}^{phys}$, $\mathbf{x}^{mind}$, $\mathbf{x}^{ctxt}$, and $\mathbf{x}^{soc}$).

The input (output) vectors represent a single individual at a particular state (physical, mental, experiential, and social). An individual that is the subject of experience or the object being referred to can be specified in identity vector $\mathbf{x}^{id}$.

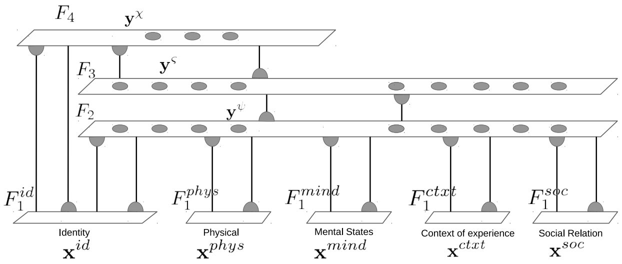


Figure 3: ARTSELF neural architecture for self-awareness

Every reflection r is presented as vectors in the input fields that initiate a resonance search to select a node in F_2. The selected node is activated (assigned with value 1) and held transiently while receiving a series of different r presentation so that a pattern of transient activations will be produced in F_2 as vector $\mathbf{y}^{\psi}$ (see Figure 4(i) at F_2). In this case, $\mathbf{y}^{\psi}$ corresponds to an introspection unit ψ_{a^*} (3) of agent a^* as defined in the previous section above.

Vector $\mathbf{y}^{\psi}$ initiates a resonance search in F_3 to select the best matching node representing an instance of introspection. The selected node in F_3 is held transiently while another pattern of introspection unit is formed in F_2 and categorized in F_3. In contrast to F_2, a sequential pattern can be formed in F_3 by gradually decaying every previous transient activation whenever a new F_3 resonance selection is conducted.

Algorithm 1: Creating a deeper level of awareness

```
if $\mathbf{y}^{\varsigma}$ does not match with a currently activated node $i$ in $F_4$
  where $y_i^{\chi} \in \mathbf{y}^{\chi}$, and ($y_i^{\chi} > 0$) then
    set $\mathbf{x}^{id}$ with the current subject $id$ ;
    select a resonance node $J$ in $F_4$ based on $\mathbf{x}^{id}$ ;
    set $y_J^{\chi} \in \mathbf{y}^{\chi}$ to 1 ;
    set every other $y_j^{\chi} \in \mathbf{y}^{\chi}$ so that
      $y_j^{\chi(\text{new})} = y_j^{\chi(\text{old})}(1 - \tau)$ ;
    Learn the sequence of $\mathbf{y}^{\varsigma}$ by associating it with node $J$ ;
end
```

The sequential pattern formation in F_3 is represented as vector $\mathbf{y}^{\varsigma}$ which corresponds to the sequence $\varsigma_{a'}$ of agent a''s introspection units (2). Let y_i^{ς} be the ith element of $\mathbf{y}^{\varsigma}$. The sequential pattern can be formed by reducing the y_i^{ς} value so that $y_j^{\varsigma(\text{new})} = y_j^{\varsigma(\text{old})}(1 - \tau^*)$ where τ^* is the decay parameter. Let y_t^{ς} be the value of an element in $\mathbf{y}^{\varsigma}$ activated or selected at time t. Over time, the gradual reduction process will form a sequential pattern that $y_t^{\varsigma} > y_{t-1}^{\varsigma} > y_{t-2}^{\varsigma} > ... > y_1^{\varsigma}$.

The top F_4 field captures the awareness χ_a in which a single node j in the field represents an ordered list of introspection units $\varsigma_{a'j}$ (2). To represent the depth of awareness level l in the hierarchy, as defined in (5), the gradual transient values are used. Algorithm 1 describes the steps taken to create a deeper level of awareness. Similar to the formation of sequential pattern in F_3 layer, the level in the hierarchy is represented as the activation value of the corresponding node in F_4 which is gradually reduced over time (see line 5 in Algorithm 1). It is ensured that no duplication or repetition of awareness can be produced in different levels (line 1 in Algorithm 1). The awareness is always associated with the subject individual in the former level of the hierarchy. Figure 4(i) shows some possible transient patterns formed in the top three layers of ARTSELF.

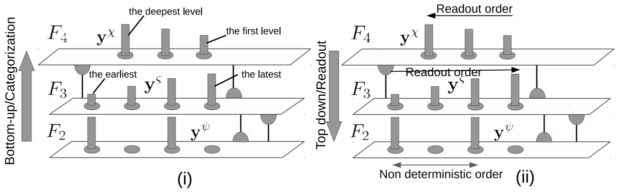


Figure 4: (i) Formation multiple, sequential, and hierarchical structure of introspection and awareness in different fields through bottom up process; (ii) Continual readout accross different layers once transient patterns are formed.

Nodes that are transiently activated in a category field will be read out to transmit their stored patterns to the fields in the lower layer. The readout follows a particular order when the activations are sequentially or hierarchically ordered. Figure 4(ii) shows different orders of readout in different layers once their nodes are activated. The readout from F_4 are conducted consecutively from the smallest (but larger than zero) to the largest node's activation as represented in vector $\mathbf{y}^{\chi}$. Similarly, nodes in F_3 layer are readout following the same order of activation values in $\mathbf{y}^{\varsigma}$.

Self as Interacting Memory Systems

ARTSELF is a model of working memory system similar to the concept of working self memory in (Conway 2005). It encodes the hierarchy of individuals including all aspects of self (identity, physical, mental, experiential, and social). The ARTSELF neural architecture is a plausible model of self-awareness that allows the intricate dynamic hierarchical-sequential structure of selves to be represented and held transiently as defined in the previous section. The neural mechanism of pattern matching enables activation and selection to be handled approximately to retrieve transient items in memory with incomplete or noisy information cues. The top-down attention and readout operations allow new patterns to be produced using the pattern completion feature of the neural network.

ARTSELF requires other systems or modules to be applicable. Following the aspects of individual above, ARTSELF can be connected with other systems that serve as repositories, models, perceptual buffer, action buffer, or executive control. For self-awareness, some external systems can be connected to the working self-memory (Figure 5): (1) *Identity catalog* provides all identities and characteristics of the individual; (2) *Physical model* provides most information on the physical aspect; (3) *Social repository* provides information about social relationship of another individual to the agent including how close the individual to the agent and also social rules or norms that the agent adopts to interact with social life; (4) *Autobiographical memory* stores every detail experiences and learns some general events and concepts.

In this paper, our focus is only on ARTSELF. The other related external memories are defined only at a high level. Operations involving external memories may not just directly retrieve information and make answers but also indicate the status of the results. When the retrieval fails or no expected results can be found, it must indicate it to the working self memory including a confidence value and/or possible options.

For example, if no matching identity can be found in the identity catalog, it will notify the working memory about the retrieval failure. It may also return an individual with similar characteristics together with some confidence value. Similarly, the autobiographical memory will notify the working memory when no exact match of memory trace can be found but similar items may be returned. In this case the working self memory can distinguish whether the agent has totally no idea about the situation or can still remember partially but unsure about the specific detail.

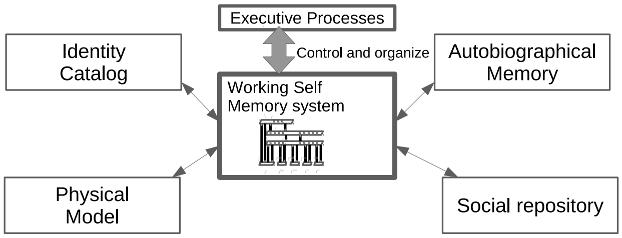


Figure 5: Agent Architecture with Self-Awareness

The main assumption about the agent architecture is that there are several modules that control memory or behavior running concurrently. Each may react independently, but they all can be triggered by changes in the working self memory system. A module can also provide some inputs to the working memory or make some changes to the existing contents in the working self memory. For example, some executive processes may remove an introspection unit about achieving a particular goal when the goal has just been achieved.

Self-Aware Conversational Robot

To demonstrate the features of ARTSELF, the self-awareness model is applied to a NAO humanoid robotic platform[1]. The robot has its own built-in modules to interact with humans. It can identify a person based on one's facial features that have been learnt before. It can talk fluently with the person using a programmable voice-recognition system. The behaviors of the robot are mainly driven by some internal modules that running concurrently. The overall architecture of self-awarenes follows the main model as depicted in Figure 5. However, each memory module also handles internal processes specific to the NAO platform.

[1] https://www.ald.softbankrobotics.com/en/cool-robots/nao

A simple scenario demonstrates self-awareness when the robot creates a representation of itself as it observes its own image in the mirror. The ability to recognize a person, and a visual image of an object (including its own) is supported by the built-in mechanism in the robot. Based on the output of the recognition system, the awareness representation is formed. Figure 6(i) shows that firstly a unit introspection is created consisting of the robot's self (identified as `me`) as the subject of observation and the robot itself as the object (as seen in the mirror). In ARTSELF, this introspection unit corresponds to the formation of vector $\mathbf{y}^{\psi}$ in F_2 layer. This vector initiates the formation of sequential pattern in F_3 and the first level of awareness in F_4 (awareness level 1). The object retained in level 1 triggers a prediction about the mind of the objective individual. In that case, the awareness formation is repeated and level 2 introspection is created following Algorithm 1 for both subjective and objective self-awareness reflecting the same individual.

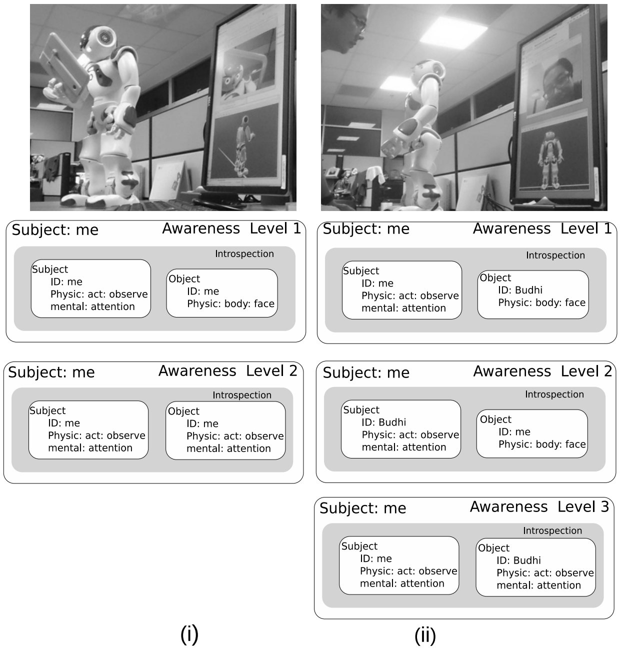
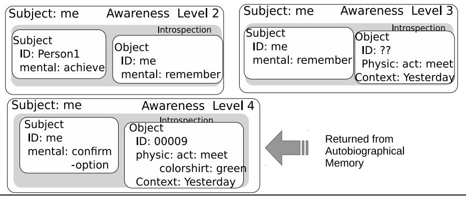
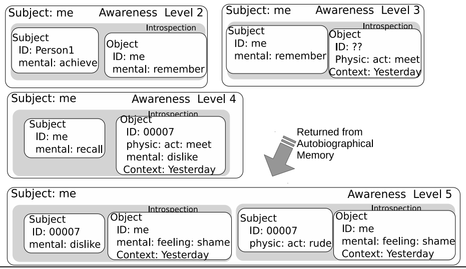
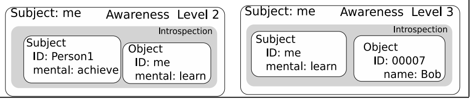


Figure 6: (i) The awareness activation in the network when the robot sees itself in the mirror; (ii) The awareness activation in the network when the robot sees and interacts with a person.

On the other hand, when the robot engages with a known person and makes eye contact (the robot has a built-in eye-gaze detection), more levels can be made as in Figure 6(ii). At level 1 of introspection, the robot, as the subject, observes the person known as '`budhi`'. As it retains in memory for some time, the robot tries to make a prediction about the other object's intentional state by creating a level 2 awareness, following Algorithm 1 in which the subject is now '`budhi`' that observes and believes about the existence of the robot's self (`me`). While it is in memory, the prediction continues to produce level 3 awareness that '`budhi`' believes (as indicated in the former level) that the robot is also observing '`budhi`'s self.

Both cases of observation by mirroring and making eye contact with another person halt the awareness to a particular level (level 2 for mirroring and level 3 for eye contact). This happens because further level of awareness will just create the same pattern as the one in the previous level. The inherent mechanism in ARTSELF (see line 1 in Algorithm 1) prevents this from happening.

Table 1: Self-Awareness in Conversation

<table>
<tr><th>Agent</th><th>Mind and Dialog Content</th></tr>
<tr><td><b>1. Person1</b></td><td>Hi, do you remember the person you met yesterday?</td></tr>
<tr><td><b>2. Robot</b></td><td>Is he the one with the green shirt that came with you?

</td></tr>
<tr><td><b>3. Person1</b></td><td>No, he is the one who doesn't like you</td></tr>
<tr><td><b>4. Robot</b></td><td>Oh, the one that's rude to me. I think I remember

</td></tr>
<tr><td><b>5. Person1</b></td><td>His name is Bob</td></tr>
<tr><td><b>6. Robot</b></td><td>Okay, his name is Bob. Got it

</td></tr>
</table>

The next case shows how the robot deals with a complex dialog involving self awareness and memory. The NAO robot is assumed to have interacted with humans and learnt about different people. Some people were deliberately rude or said bad things to the robot while it learned all the experiences in autobiographical memory. The next scenario reveals the role of self-awareness in aligning the conversation towards mutual understanding. The robot asks for confirmation about possible remembered objects of discourse so that the interlocutor can refer to the same topic or context.

Table 1 shows a dialog excerpt between **person1** and the **robot**. From the first question raised in the first row, the robot produces a response according to the incomplete cue in the request to remember a yesterday's event (row 2). The table also reveals the internal representation of self-awareness in ARTSELF when the **robot** starts to produce a response (just below the dialog content of the **robot**). To deal with the incomplete inquiry (Awareness Level 3 with unknown object ID in row 2 of the robot's responses), the **robot** selects one possible option as it is returned by the robot's autobiographical memory and asks the user for confirmation (Awareness Level 4 in row 2). A clue from **person1** (Awareness Level 4 in row 4) makes the **robot** represent itself as the object in the introspection in which the subject is an unknown person (Awareness Level 5 in row 4). Once the target is captured in one's mind, the next instruction enables the agent to learn directly from the person (Awareness level 3 in row 6).

This later scenario demonstrates the formation of objective self-awareness that an agent views itself from a third person.

Conclusion

A neural network model of self-awareness called ARTSELF has been presented. By including identity as the main characteristic of an individual in addition to physical, mental, experiential, and social aspects, a complex structure of possible self hierarchy can be formed to support self-awareness. The model is demonstrated to support both subjective and objective self-awareness. Using the NAO robotic platform, it is demonstrated that the model allows a human to have conversation and asking about the robot's social life and opinions as if the robot were an individual with personal identity, experiences, and social life.

In the future, the study will be extended to explore more complex interactions. Besides responding to questions, the agent can be made to display more complex emotions related to self such as embarrassment, guilt, ashame, pride and so on at the appropriate occassion to add the realism of the model.

Acknowledgements

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative and Singapore Ministry of Education Academic Research Fund (AcRF) Tier-1 (RG137/14).

References

Bongard, J.; Zykov, V.; and Lipson, H. 2006. Resilient machines through continuous self-modeling. *Science* 314(5802):1118–1121.

Bringsjord, S.; Licato, J.; Govindarajulu, N. S.; Ghosh, R.; and Sen, A. 2015. Real robots that pass human tests of self-consciousness. In *Proceedings of the Twenty-Fourth IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)*, 498–504.

Conway, M. A. 2005. Memory and the self. *Journal of Memory and Language* 53(2005):594–628.

Duval, S., and Wicklund, R. A. 1972. *A Theory of Objective Self Awareness*. New York: Academic Press.

Grossberg, S. 2013. Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks* 37(2013):1–47.

Hart, J. W., and Scasselati, B. 2012. Mirror perspective-taking with a humanoid robot. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2012)*, 1990–1996.

Lewis, P. R.; Chandra, A.; Parsons, S.; Robinson, E.; Glette, K.; Bahsoon, R.; Torresen, J.; and Yao, X. 2011. A survey of self-awareness and its application in computing systems. In *Proceedings of the Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops (SASOW 2011)*, 102–107.

Neisser, U. 1997. The roots of self-knowledge: Perceiving self, it, and thou. *Annals of the New York Academy of Sciences* 818:19–33.

Pfeifer, J. H.; Lieberman, M. D.; and Dapretto, M. 2007. "I know you area but what am I?!": Neural bases of self- and social knowledge retrieval in children and adults. *Journal of Cognitive Neuroscience* 19(8):1323–1337.

Ramachandran, V. S. 2011. *The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human*. New York: W.W. Norton & Company.

Samsonovich, A. V., and Nadel, L. 2005. Fundamental principles and mechanisms of the conscious self. *Cortex* 41(5):669–689.

Sloman, A., and Chrisley, R. 2003. Virtual machines and consciousness. *Journal of Consciousness Studies* 10(4–5):133–172.

Stoytchev, A. 2011. Self-detection in robots: a method based on detecting temporal contingencies. *Robotica* 29(1):1–21.

Subagdja, B., and Tan, A.-H. 2012. iFALCON: A neural archiecture for hierarchical planning. *Neurocomputing* 86:124–139.

Subagdja, B., and Tan, A.-H. 2015. Neural modeling of sequential inferences and learning over episodic memory. *Neurocomputing* 161(2015):229–242.

Sun, R. 2007. The importance of cognitive architecture: An analysis based on CLARION. *Journal of Experimental and Theoretical Artificial Intelligence* 19(2):159–193.

Tan, A.-H.; Carpenter, G.A.; and Grossberg, S. 2007. Intelligence Through Interaction: Towards A Unified Theory for Learning. In *International Symposium on Neural Networks (ISNN) 2007*, LNCS 4491, 1098–1107.

Wang, D.; Tan, A.-H.; and Miao, C.-Y. 2016. Modeling autobiographical memory in human-like autonomous agents. In *Proceedings of the Fifteenth International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2016)*, 845–853.

Wang, D.; Subagdja, B.; Tan, A.-H.; and Ng, G.-W. 2009. Creating human-like autonomous players in real-time first person shooter computer games. In *Proceedings of the Twenty-First Annual Conference on Innovative Applications of Artificial Intelligence (IAAI'09)*, 173–178.

Wang, W.; Subagdja, B.; Tan, A.-H.; and Starzyk, J. A. 2012. Neural modeling of episodic memory: Encoding, retrieval, and forgetting. *IEEE Transactions on Neural Networks and Learning Systems* 23(10):1574–1586.