

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

1-2021

Who am I?: Towards social self-awareness for intelligent agents

Budhitama SUBAGDJA

Singapore Management University, budhitamas@smu.edu.sg

Han Yi TAY

Ah-hwee TAN

Singapore Management University, ahtan@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research



Part of the [Artificial Intelligence and Robotics Commons](#)

Citation

1

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylids@smu.edu.sg.

Who Am I?: Towards Social Self-Awareness for Intelligent Agents

Budhitama Subagdja¹, Han Yi Tay² and Ah-Hwee Tan^{2,3}

¹Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly,

²School of Computer Science and Engineering,
Nanyang Technological University, Singapore

³School of Information Systems, Singapore Management University
budhitama@ntu.edu.sg, c160103@e.ntu.edu.sg, ahtan@smu.edu.sg

Abstract

Most of today's AI technologies are geared towards mastering specific tasks performance through learning from a huge volume of data. However, less attention has still been given to make the AI understand its own purposes or be responsible socially. In this paper, a new model of agent is presented with the capacity to represent itself as a distinct individual with identity, a mind of its own, unique experiences, and social lives. In this way, the agent can interact with its surroundings and other agents seamlessly and meaningfully. A practical framework for developing an agent architecture with this model of self and self-awareness is proposed allowing self to be ascribed to an existing intelligent agent architecture in general to enable its social ability, interactivity, and co-presence with others. Possible applications are discussed with some exemplifying cases based on an implementation of a conversational agent.

1 Introduction

In the field of AI, an intelligent agent is an autonomous entity that interacts reactively with the environment, socially with other agents, and proactively in a goal-directed manner [Wooldridge and Jennings, 1995]. More and more agents and autonomous machines may work alongside humans in the near future. As the agents get much closer to humans in their work and operation, they need to be aware of their surroundings especially people within the proximity. It is crucial to ensure that the co-existing agents interact safely with humans while still taking benefit as much as possible from their collaboration. In this case, how should agents be developed to ensure safety, trustableness, and engagingness should be paid more attention.

In this paper, the notion of self is seen to be the key aspect that enables an AI agent to have a genuine understanding about its surroundings. Appreciating what appears in the world or understanding why others do in a certain way constitutes the context regarding the agent itself as the one that conducts the observation and/or shares the domain environment. One illustrating example has been outlined in a classic science fiction story 'Runaround' [Asimov, 1942] which is

famously known with 'three rules of robotics' that an agent (robot) shall not harm a human being, shall obey orders from humans, and shall protect its own existence as long as it does not compromise the other rules. Regardless the inapplicability of the laws to govern the robots today, interpreting the rules like those requires the agent at least to be able to comprehend that they are referring to its own self as the subject of the laws in the first place. At any time, an intelligent agent should be able to know that itself is a unique individual like others (e.g. humans) that can also think or know about them. As itself, the agent can behave as an individual, be a part of a group of co-existing others, make an individual contribution with them, or just leave away. Self helps one to cope with one's social world. It also glues together different parts of the individual so that they can work together coherently as one. All in all, the self is necessary for an individual to bridge the internal dynamic of one's mind with the social world and making sense of them.

Despite many applications of interactive agent like social robots, personal assistants, or robotic companion [Biundo *et al.*, 2016; O'Brien, 2018] have been around today to entertain, to assist people, and even to develop social relationship [Terdiman, 2018; Druga *et al.*, 2017; Satariano *et al.*, 2018], no real competence of understanding oneself nor the others' exist in the agent's mind. Current AI programs and agents learn and acquire knowledge from a huge amount of data without getting the meaning for themselves. The underlying principles, methods, and frameworks to make an agent self-aware are still lacking.

Inspired by the concept of self in cognitive science [Neisser, 1993] and social psychology [Hegtveldt and Johnson, 2018], this paper presents a new model of self-awareness for intelligent agents that interact with the world or other agents (e.g. humans). It is proposed that self-awareness can be achieved whenever the agent maintains a model or a representation about itself in its own mind in the same way as representing the others. Ascribed with self, an agent becomes a unique individual with identity, mind of its own, memory of experiences, and social life as envisaged in [Subagdja and Tan, 2019]. The self-awareness consists of some representation of individuals in the agent's mind including the agent itself as subject or object associated with properties or relations. In this way, the awareness of self becomes dynamic and can be associated with other kinds of domain specific model

and representation. This paper also proposes a framework for developing a self-aware agent wherein a domain specific operations and functionalities can be enhanced with the model to make a self-aware social agent.

This paper is organized as follows. Section 2 discusses related work. Section 3 conceptualizes self and self-awareness for computational agents. Section 4 describes the proposed architecture and framework for the self-aware agent in more detail. Section 5 exemplifies the the model and the framework in a therapeutic conversational agent application. Section 6 concludes the paper and discusses future work.

2 Related Work

Some works have been done in modeling self-awareness computationally. Most of them consider self-awareness as a form of reflection or introspection of their own on-going computational process so that their reasoning, planning, learning, or meta-cognition can be improved [Cox, 2007; Bringsjord *et al.*, 2015; Sloman and Chrisley, 2003; Sun, 2007; Samsonovich and Nadel, 2005]. In this introspective model, an AI system or program can adapt its own computational state and optimize the process on-the-fly. These works, however, do not look at the identity as well as social aspects of self wherein the self may influence other agents and vice versa. These model concern only to the internal aspect of self to optimize the system.

The reflective computational model has also been generalized to include the external or social dimension as an engineering principle for developing a computing system [Lewis *et al.*, 2016]. It is used to specify what information should be determined as artifacts for developing the system as if it is an individual with a cognitive system. This engineering approach is then further developed for characterizing systems with self-awareness in general [Kounev *et al.*, 2017]. Here, it is viewed that a self-aware system is a system that learns a model and captures knowledge about itself and its environment in an ongoing basis. The system can then use the learned model about itself to reason and act based on it. This general definition for self-aware computing systems [Lewis *et al.*, 2016; Kounev *et al.*, 2017], however, does not really specify what or who should be accounted as the individual self to be modeled nor how to distinguish oneself from another. There is still no guideline to demarcate the things to be modeled as parts of oneself from the things belong to others in the first place.

Other works on self-awareness have focused more on modeling limited aspects of self to make the agent recognize and identify itself as a distinct entity from others. Works on embodied robotic agents focus on the discovery of a robot's own motion dynamic [Bongard *et al.*, 2006; Hart and Scasselati, 2012; Stoytchev, 2011], recognize and analyze oneself based on mirror image projection [Hart and Scasselati, 2012; Stoytchev, 2011], or control and adapt its own movement robustly despite impairments [Bongard *et al.*, 2006]. With some representation of second order self-awareness and theory of mind, an agent can anticipate the behavior of other agents in interactive tasks including conversation and story telling [Winfield, 2018; Chatila *et al.*, 2018]. Similar to this

approach suggests anticipation through internal mental simulation [Gray and Breazeal, 2014]. A more complex model for anticipating self and others in an arbitrary number of level or order of self-awareness has also been proposed in [Subagdja and Tan, 2017] with a neurally-plausible model of working memory. In contrast, the proposed model in this paper only consider subjective and objective level of self-awareness.

Following the envisioned concept of self-awareness in [Subagdja and Tan, 2019], the proposed model in this paper also adopts multi-level perspectives to anticipate possible futures of other individuals. In contrast to the other models mentioned in this section, it emphasizes identity as the main distinctive aspect to consider besides mind, social relationships, and autobiographical memory [Wang *et al.*, 2019] that allows consistent references to the same individual over different times and possible contexts. However, the ecological aspect, as suggested in [Neisser, 1993], is adopted in place of embodiment to include the condition of one's environment.

3 Self and Self-Awareness

Self-awareness is known to be the capacity to put oneself as the focus of attention [Duval and Wicklund, 1972]. In this section, the awareness structure of an agent will firstly be defined before conceptualizing the notion of self and self-awareness as the main topic in this paper.

3.1 Awareness Structure and Perspectives

An agent can be aware of something if the agent maintains a model about the thing in memory as a result of or as an input to a particular computational process. The model is a simplified description or a quick-sketch of the actual thing being represented. Here, identifying an object also means that the object must be unique. When the agent pays attention to an object and identifies it as a box, for example, the existence of the box should be unique. Similarly, when it perceives an agent (e.g. person), it must refer to a unique individual or person.

Attention unit $\bar{a} = \{p_1, \dots, p_2\}$ is defined as a representation or model of something created in mind when the agent focuses its attention to something, an object, or an individual. $p_i \in \bar{a}$ is an attribute that describes \bar{a} . \bar{a} may also refer to a relation between attention units. For example, "distance" is a relation between two objects in space with a value indicating the spatial distance between the two. *Awareness structure* $\bar{\mathcal{A}} := \{\bar{a}_1, \dots, \bar{a}_n\}$ consists of all attention units the agent has at a particular moment. Figure 1a illustrates the contents of a basic awareness structure of an agent when the agent perceives or pays attention to every object in the environment (box, cylinder, and human). Relations among the object may also be represented as attention units (e.g. distances among the objects).

Figure 2a shows an example of the content of the awareness structure in a graph representation when the agent is at situation like in Figure 1a. The agent has two objects in mind (box_00 and cylinder_00) each as Object typed object and one individual agent (agent_00 as Agent typed object). The agent also maintains relational attentions between objects (distances). Once an attention unit is put in the awareness structure, more information may be filled up to complete

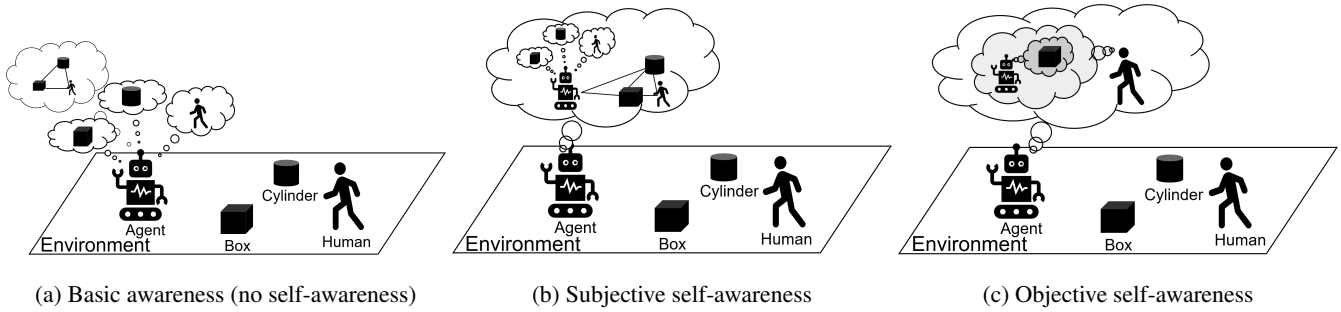


Figure 1: Different levels of self-awareness

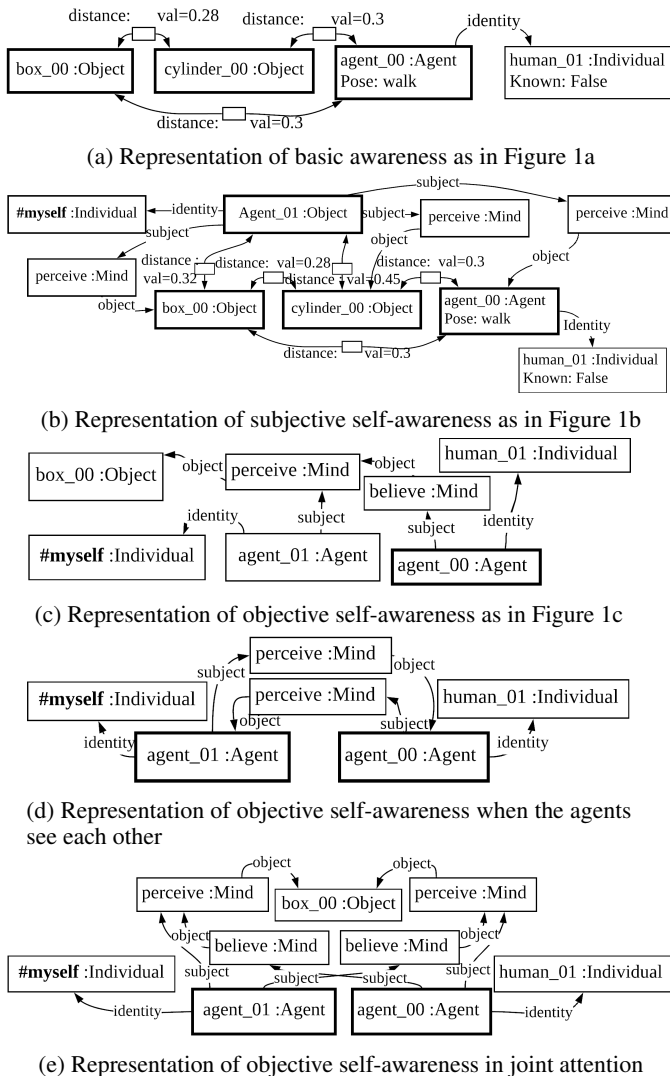


Figure 2: Corresponding awareness representation of Figure 1.

the representation. For example, `agent_00` is characterized with `Pose:walk` and identified as `Individual` with identity `human_01` who is also unknown or a stranger. The attention units of object (or agent) with thick lines indicate that they are in *subjective* level of awareness as objects being observed first-hand by the agent.

3.2 The Concept of Self and Self-Awareness

Agent is a special type of object in the awareness structure since it is always considered to be an individual associated with a unique identity. As a unique individual, it can be characterized with five aspects adapted from [Subagdja and Tan, 2019] as defined in Table 1. Whenever an agent is represented in the awareness structure, some computation process will be initiated to characterize the individual based on the aspects defined in the table. For example, when an attention unit is identified as `Agent`, it will be associated with an identity of a particular individual (e.g. `human_01`). Other attributes related to the ecological aspect [Neisser, 1993] may follow to be filled (e.g. “walk” body pose, location, distances). The characterization process may continue to identify the particular social relationship with the person (social aspect) by consulting the memory. In case it is a first encounter with the person, the attention unit may be marked as unknown or stranger.

In terms of identity, an individual can also be characterized based on its type. When someone is known (unknown) or has a (no) particular social relation with the subject, then the one is identified as “Known” (“Stranger”) individual. One can also be a “Generalized others” referring to anyone belongs to a social group or a set of individuals that share the same characteristics regardless who is the individual in particular. This “generalized” identity has been conceptualized with the same term in social psychology [Mead, 2015] to explain that one should have a common expectation about actions and thoughts of anyone within a society or a social group. Finally, “Oneself” refers to a unique individual that defines the agent itself as the subject and as a single individual.

Self-Awareness occurs whenever an individual (`Agent` typed attention unit), with “Oneself” identity, is present in the agent’s mind. There are two kinds of self-awareness:

- *Subjective self-awareness*. There is a representation of oneself in the subjective level of awareness wherein the agent thinks or “feels” about it first-hand as illustrated in Figure 1b. Figure 2b shows the corresponding representation in the awareness structure.

- *Objective self-awareness.* There is a representation of oneself as an object of another individual’s mind within the awareness structure of the agent as illustrated in Figure 1c. Figure 2c shows the corresponding representation in the awareness structure.

Figure 2b and 2c show that the self-awareness is indicated as a reference to **#myself** indicating a “Oneself” identity. A particular type of relation is also exemplified in the figures. A relation with `Mind` type indicates a mental state of an individual (“Mind” aspect) which is also a kind of attention unit but can be related to other attention units as subject or object. In this case, `perceive` and `believe` are two relations as parts of the “Mind” aspect of the individual.

In the case of subjective self-awareness (Figure 2b), the agent represents itself (as `Agent_01`) at subjective level as a subject of relation (`perceive:Mind`) to an object (`box_00`, `cylinder_00`) or another individual (`agent_00`). In objective self-awareness (Figure 2c), the agent itself (`Agent_01`) becomes a part of the object of a relation (`believe:Mind`) in another individual’s mind (`agent_00`). In another interesting case, the agent may also have both subjective and objective self-awareness at the same time. When the two agents look at each other or both are engaged in a dialog, one can be the subject of a relation to another and the object of a relation from the other both at the subjective level of awareness. Figure 2d shows the representation wherein the two different individuals look at (`perceive`) each other. A similar condition can occur when both have a joint attention to an object like in Figure 2e. This kind of reciprocal relationships may occur when the agents are actively engaged in a conversation.

Figure 3 shows a simplified data model of all attention units in the awareness structure covering all aspects of the individual. It is shown that some of relational and attribution of an attention unit can have a value as a property to indicate the strength or magnitude (positive or negative) of the relation or the attribute. They may also be associated with a confidence level representing its (un)certainty with positive or negative likeliness. Every aspect is made to be generic and flexible with relations to its attribute that can be added incrementally (e.g. `RelAttribute`, `GroupAttribute`). Event object takes all attention units in the awareness structure at a certain moment, as represented in its `timestamp` attribute, as a snapshot. This snapshot is then recorded in memory as a trace of the awareness in episodic memory (`episode:Memory`). Later on, some items recorded may be retrieved back to

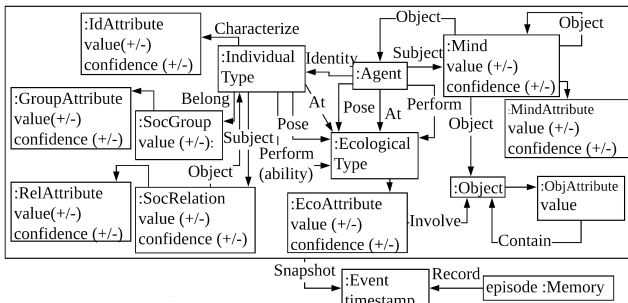


Figure 3: Simplified Data Model of the Awareness Structure.

Algorithm 1 Executive Process Operation Cycle

```

1: while true do
2:   for every object  $o$  produced by PS and every execution
      status update  $s^a$  returned by AES do
3:     create attention unit  $\bar{a} \leftarrow f^p(o)$  to update AS.
4:     update or remove attention unit in AS accordingly.
5:   end for
6:   for every  $\bar{a}^s$  in AS do
7:     if  $\bar{a}^s$  is an intention of oneself at subjective level
      then
8:       add plan or action type  $\bar{a}^{sa}$  to realize  $\bar{a}^s$ .
9:     else if  $\bar{a}^s$  is an action type at subjective level then
10:      create action description  $a \leftarrow f^d(\bar{a}^s)$ .
11:      send  $a$  to AES for execution.
12:     else
13:       update AS based on domain knowledge or norms.
14:     end if
15:   update AS based on a generalized model of individ-
      uals.
16: end for
17: end while
    
```

awareness structure when needed by a later reasoning process.

4 Self-Aware Agent Architecture

A general architecture of self-aware agent can be made according to the conceptual model described in the previous section with a structure shown in Figure 4. Here, Perceptual System or PS is a component of the architecture that continuously perceives the environment to identify an object o as a representation of the thing being perceived. PS may also receive some parameters to adjust its focus of attention or its filter of information to refine the object identification. As a significant part of the architecture, the Executive Process (EP) controls and regulates the production and use of Awareness Structure (AS) in Working Memory (WM). Through Interaction Manager (IM), the object representation o is interpreted with function $f^p(o)$ to produce an attention unit \bar{a} about the object o to update AS. Interaction Manager (IM) also scan the content of AS to find any agent’s own intention applicable to execute and then update its status or generate an action description a with function $f^d(\bar{a})$ to be executed or performed by Action Execution System (AES). Besides IM, Reasoning and Inference Engine (RIE) work also to update AS based on the current content of AS and/or entries in Autobiographical Memory (AM). For every Agent type attention unit in subjective level of AS, RIE may also make some predictions or expectations about what in the agent’s mind based on a model of “Generalized others” individual or the common behavior of people within the same social group with the agent. Periodically, RIE goes through every attention unit to check for consistency and to purge the unit into AM more permanently as a memory consolidation. AM is an information storage for events experienced by the individual [Wang *et al.*, 2019] with the same structure as data model in Figure 3 but retain the information more permanently. AM may also store general-

Aspect	Description
Identity	a set of properties and relations to identify the agent as a unique individual
Ecological	a set of properties and relations defining the individual in relation to its environment, embodiment, and its own capabilities [Neisser, 1993]
Mind	a set of properties and relations indicating the mental states of the individual at one moment that may point to another object, individual, or relation as the object of the mental operation
Memory	consisting of traces and snapshots of awareness structure from time to time
Social	a set of other individuals, their relationship with the agent itself, and social groups wherein other agents or itself may belong to

Table 1: Aspects of Individual

ized characteristics of people from the same social group or cluster to be used for later prediction. Algorithm 1 shows the overall operation cycle of EP to maintain the consistency of AS.

The architecture describe above can be considered as a domain generic framework for developing an agent with self-awareness. In Algorithm 1 line 13, For instance, the knowledge and social norms for updating the attention unit are still left unspecified. This allows the architecture to be customized to base on some domain specific knowledge or social norms.

5 Application and Examples

The self-aware agent model has been implemented as an agent application to carry out Individual Cognitive Stimulation Therapy or ICST as a psycho-social non-pharmacological intervention to enhance cognition for an individual suffering from dementia or cognitive deficit through one-to-one interactions and activities with a carer involving mental stimulation, reminiscence, and reality orientation [Ali *et al.*, 2018]. The agent is made conversational as a chatbot to complement the human carer. In particular, it conducts a number of activity sessions in the therapy. The design of the ICST agent follows the generic architecture in Figure 4 customized to cater the specific needs (Figure 5). The agent receives inputs from the user and produces the response output as natural language utterances or texts. A natural language parser interprets the text from the user input into a dialog intent as a structure that specifies what the user intends by that text. Similarly, a natural language generation module processes a communication act that specifies the chatbot’s intention into natural language text. The Dialog Manager (DM) controls the flow of the dialog based on the awareness structure and ICST knowledge base. Self-awareness in this case is used to handle two kinds of strategy of conversation. The first one is for adapting the dialog based on the user feeling

towards the agent’s intention. The excerpt below shows a case of the ICST session, where the bot tries to stimulate the user’s cognition through reminding the person about the situations in daily lives.

```

Eunice (Bot) : It is so wonderful to be able to see you again Budhi
                ... How is the weather like today?
Budhi (User) : The weather today is sunny
                ...
                ...
Budhi (User) : I was hanging out with family this week
Eunice (Bot) : Wow, it’s sounds interesting!
                I’ve read the news lately
                I’ve heard PMDs (Personal Mobility Devices) have
                been banned this week also
                what do you think about it?
Budhi (User) : What! not again. I don’t want to talk about it
Eunice (Bot) : Oh, sorry!
                I thought you might be interested
                Let’s talk about something else
                How about your time with family in the past week?
Budhi (User) : It’s great! I was having fun
Eunice (Bot) : good. So what did you do ?
                ...
    
```

The bot applied a strategy to keep stimulating the user’s cognition by asking the user about various issues regarding daily lives. Figure 6 shows the goals to maintain the *dialog engagement* and *joint attention* in the related topic (see also Figure 2d and 2e) of PMDs ban from the news. However, at a certain point the user responded negatively with “...I don’t want to talk about it” indicating dislikeness to the topic. In this case, the agent identified negative interest towards the topic (negative interest :Mind) and low or negative sentiment towards the agent.

The self-awareness model allows the agent to keep track of not just the user’s intentions but also its own. By consulting its own awareness structure to identify a break in the *joint attention*, the agent can pinpoint which object or attention unit the user has a negative intention so that it can

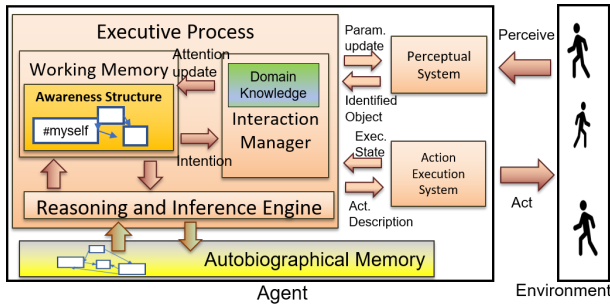


Figure 4: An Agent Architecture with Self-Awareness.

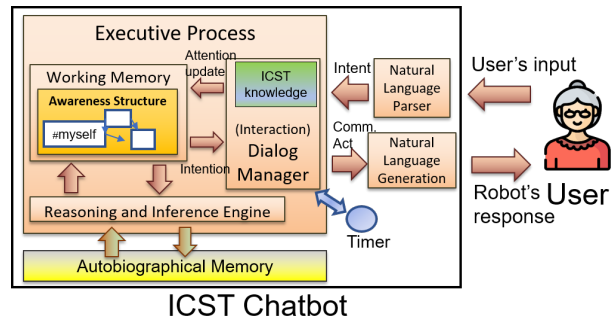


Figure 5: Architecture of a Conversational Agent with Self-Awareness for ICST.

fix the situation by signaling the user to switch the attention to another topic (“Let’s talk about something else”). The second strategy is *self-disclosure* to support stimulating the user cognition by revealing information about oneself to another so that the other person may also be willing to self-disclose in turn [Altman and Taylor, 1973]. This strategy is used in this ICST chatbot to stimulate the user to be engaged in the dialog by initially self-disclosing as in the following excerpt from an ICST activity session with the chatbot.

```

...
Eunice (Bot) : Okay! let’s get on with our activity
                Today, we will be doing an activity related to food as
                well
Budhi (User) : so what would we be doing ?
Eunice (Bot) : ... we will be listing foods that have the same first letter!
                For example, for the letter C there is carrot cake, chicken
                sandwich, chesse pizza and corn dog!
...
                Let’s start with the letter A. List four food with the first
                letter A!
Budhi (User) : [long silence]
Eunice (Bot) : come on! can you do it?
Budhi (User) : Yes, I’m still thinking..
Eunice (Bot) : okay, I give you a clue, Almonds
                what else? hmm I also can’t remember so well
Budhi (User) : okay, maybe Apple pies ?
Eunice (Bot) : that’s good
                I think you better than me on this. okay, what else?
Budhi (User) : Arugula, Alfafa sprouts
Eunice (Bot) : that’s great
                Actually, I just knew, there are foods with those
                names.
                Okay, let’s try another letter. B! List four food with
                the first letter B!
Budhi (User) : Okay! Broccoli, Beef Burger, Baklava..
...
    
```

In this second case, the bot identifies a problem when there is a long silence. As shown in Figure 5, the pause can be identified by Dialog Manager with the Timer module. The silence can be caused by confusion as the task may be too hard for the user. To handle the situation, the agent initiates the self-disclosure strategy by trying to perform the task itself to give some clues to the user. During this process the agent is self-disclosing its own private view and evaluation about its own capability or feelings (“.. I also can’t remember so well”, “I think you better than me on this”, “Actually I just knew, there are foods with those names”). By revealing this internal private information to the user as if it is from the subjective level of awareness of the agent, it may help to bring back the engagement with the user following the self-disclosure strategy. In fact, this strategy of self-disclosure is applied in most of the activity sessions in ICST like in the first example. Some dialog excerpt taken from the first dialog shows that self-disclosure is used to repair the dialog so that the user can still be engaged in conversation.

```

...
                what do you think about it?
Budhi (User) : Hey, I don’t want to talk about it
Eunice (Bot) : Oh, sorry!
                I thought you might be interested
                Let’s talk about something else
                How about your time with family in the past week?
...
    
```

After identifying that the issue is related to what it previously said, the bot makes the repair by revealing its internal state about the condition (“I thought you might be interested”).

Overall, the exemplified cases above requires the agent to keep track not just the information about the user but also

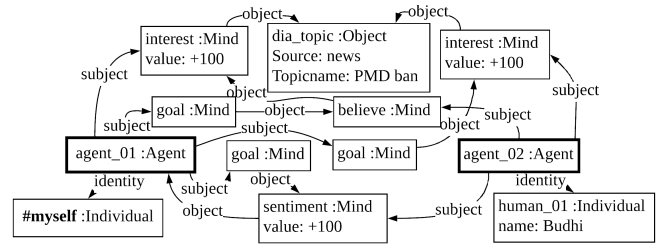


Figure 6: Awareness Structure of the ICST agent.

about its own self. In a dialog engagement like in the ICST sessions, what the user thinks or feels may likely relate to the agent’s intentions during the dialog. In the current stage of development, the awareness structure is still covering a limited aspects of an individual. Most responses from the agent are produced and controlled by the ICST manager. More dynamic and lively responses from the agents can be generated when most aspects are covered.

6 Conclusion

In this paper, it is suggested that to make an intelligent agent social and co-exist with people, the agent needs to know and understand its own self besides knowing the others. Accordingly, a model and framework of self and self-awareness for an intelligent agent has been proposed which emphasizes the capacity of the agent to see itself in its entirety as a unique individual with identity, mind, memory, and social life of its own. Awareness of self is considered achieved whenever the agent maintains a model or representation about an aspect of individual as mentioned associated with a reference to its own unique identity to support its own on-going computation process. It has been demonstrated in exemplified cases of the implemented model that the agent can adapt its own way of interacting and use self disclosure strategies to stimulate and initiate the dialog engagement.

Apart from its potential, the proposed model and framework still deserve more study and investigation. A study involving humans that interact with the agent enhanced with self-awareness should be conducted to evaluate the model and to ensure that if it is really useful. The inference mechanisms and operations to update and populate the awareness structure can still be improved or enhanced to be more realistic and human-like. This later point may also open up the possibility of using the model to study human psychological development of self and personality through simulation and modeling.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003) and, in part, by Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

References

- [Ali *et al.*, 2018] Afia Ali, Emma Brown, Aimee Spector, Elisa Aguirre, and Angela Hassiotis. Individual cognitive stimulation therapy for people with intellectual disability and dementia: protocol of a feasibility randomised controlled trial. *BMJ Open*, 8(12), 2018.
- [Altman and Taylor, 1973] Irwin Altman and Dalmas A. Taylor. *Social Penetration: the development of interpersonal relationships*. Holt, New York, 1973.
- [Asimov, 1942] Isaac Asimov. Runaround. *Astounding Science Fiction*, 29(1):94–103, 1942.
- [Biundo *et al.*, 2016] Susanne Biundo, Daniel Höller, Bernd Schattenberg, and Pascal Bercher. Companion-technology: An overview. *Künstliche Intelligenz*, 30(1):11–20, 2016.
- [Bongard *et al.*, 2006] Josh Bongard, Victor Zykov, and Hod Lipson. Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121, 2006.
- [Bringsjord *et al.*, 2015] Selmer Bringsjord, John Licato, Naveen Sundar Govindarajulu, Rikhiya Ghosh, and Atriya Sen. 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)*, pages 498–504, 2015.
- [Chatila *et al.*, 2018] Raja Chatila, Erwan Renaudo, Mihai Andries, Ricardo-Omar Chavez-Garcia, Pierre Luce-Vayrac, Raphael Gottstein, Rachid Alami, Aurélie Clodic, Sandra Devin, Benoît Girard, and Mehdi Khamassi. Toward self-aware robots. *Frontiers in Robotics and AI*, 5(88), 2018.
- [Cox, 2007] Michael T. Cox. Perpetual self-aware cognitive agents. *AI Magazine*, 28(1):32–45, 2007.
- [Druga *et al.*, 2017] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. ”hey google is it ok if i eat you?”: Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children (IDC’17)*, pages 595–600, 2017.
- [Duval and Wicklund, 1972] Shelley Duval and Robert A. Wicklund. *A Theory of Objective Self Awareness*. Academic Press, New York, 1972.
- [Gray and Breazeal, 2014] Jesse Gray and Cynthia Breazeal. Manipulating mental states through physical action. *International Journal of Social Robotics*, 6:315–327, 2014.
- [Hart and Scasselati, 2012] Justin W. Hart and Brian Scasselati. Mirror perspective-taking with a humanoid robot. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2012)*, pages 1990–1996, 2012.
- [Hegtvedt and Johnson, 2018] Karen A. Hegtvedt and Cathryn Johnson. *Social Psychology: Individuals, Interaction, and Inequality*. SAGE, Thousand Oaks, 2018.
- [Kounev *et al.*, 2017] Samuel Kounev, Jeffrey O. Kephart, Aleksandar Milenkoski, and Xiaoyun Zhu, editors. *Self-Awareness Computing Systems*. Springer, Cham, 2017.
- [Lewis *et al.*, 2016] Peter R. Lewis, Marco Platzner, Bernhard Rinner, Jim Tørresen, and Xin Yao, editors. *Self-Aware Computing Systems: An Engineering Approach*. Springer, 2016.
- [Mead, 2015] George Herber Mead. *Mind, Self, and Society: the definitive edition*. University of Chicago Press, 2015.
- [Neisser, 1993] Ulric Neisser, editor. *The perceived self: Ecological and interpersonal sources of self-knowledge*. Cambridge University Press, Cambridge, 1993.
- [O’Brien, 2018] Matt O’Brien. Robots are getting more social. are humans ready? *Associated Press*, August 9, 2018.
- [Samsonovich and Nadel, 2005] Alexei V. Samsonovich and Lynn Nadel. Fundamental principles and mechanisms of the conscious self. *Cortex*, 41(5):669–689, 2005.
- [Satariano *et al.*, 2018] Adam Satariano, Elian Peltier, and Dmitry Kostyukov. Meet zora, the robot caregiver. *The New York Times*, November 23, 2018.
- [Sloman and Chrisley, 2003] Aaron Sloman and Ron Chrisley. Virtual machines and consciousness. *Journal of Consciousness Studies*, 10(4–5):133–172, 2003.
- [Stoytchev, 2011] Alexander Stoytchev. Self-detection in robots: a method based on detecting temporal contingencies. *Robotica*, 29(1):1–21, 2011.
- [Subagdja and Tan, 2017] Budhitama Subagdja and Ah-Hwee Tan. Towards a brain inspired model of self-awareness for sociable agents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 4452–4458, 2017.
- [Subagdja and Tan, 2019] Budhitama Subagdja and Ah-Hwee Tan. Beyond autonomy: The self and life of social agents. In *Proceedings of the Eighteenth International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)*, pages 1654–1658, 2019.
- [Sun, 2007] Ron Sun. The importance of cognitive architecture: An analysis based on clarion. *Journal of Experimental and Theoretical Artificial Intelligence*, 19(2):159–193, 2007.
- [Terdiman, 2018] Daniel Terdiman. Here’s how people say google home and alexa impact their lives. *Fast Company*, 01.05.18, 2018.
- [Wang *et al.*, 2019] Di Wang, Ah-Hwee Tan, Chunyan Miao, and Ahmed A. Moustafa. Modelling autobiographical memory loss across life span. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 1368–1375, 2019.
- [Winfield, 2018] Alan F. T. Winfield. Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, 5(75), 2018.
- [Wooldridge and Jennings, 1995] Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.