Towards a Logic-based Analysis and Simulation of the Mirror Test

Naveen Sundar Govindarajulu
Rensselaer AI & Reasoning Lab
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
govinn@rpi.edu

ABSTRACT
In this paper, we examine the mirror test for self-consciousness. In the mirror test, an animal is anesthetized and a red splotch is placed on its forehead; then it is woken up and placed in front of a mirror: the animal passes the test if it removes the red splotch. Our goal is twofold: 1) to formally analyze the mirror test; and 2) to construct an artificial logic-based agent capable of passing this test. Our formal analysis and engineering is based on the Cognitive Event Calculus (CEC). We present a simple agent specification formalism, AS, based on the CEC. For an artificial logic-based agent to pass this test, it is necessary for the agent to be capable of expressing de se (of self) beliefs separate from de dicto (of word) and de re (of object) beliefs. Towards this end, we present a simple modification of the (CEC) to handle de se beliefs in a fashion paralleling Castañeda’s suggestions. The main contributions of this paper are 1) AS: a simple and formal specification of a CEC-based agent system; and 2) a partial formal analysis of the mirror test.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

General Terms
CEC, Mirror Test, De se beliefs and knowledge

Keywords
2011 EASSS, CEC, Mirror Test

1. INTRODUCTION
The purpose of this paper is to present and discuss a partial formal analysis and simulation of the mirror test for self-consciousness used in animal behaviour studies. We are in the process of achieving this by constructing an agent that is capable of passing the mirror test; the agent is based on a simple agent specification introduced here1. While the test is prominent in studies of animal self-consciousness, there is, at present, no formal analysis of the assumptions used in the test, its hypotheses etc. This is the main contribution of this project. While our formal analysis is neither complete nor final and is work in progress, we hope to work on the formalization with feedback from the mirror testing community and logic-based agent modeling community. In order to carry out the formal analysis of the test, we need a suitable logical calculus that is rich enough to model intensionalities (beliefs, knowledge, desires, intentions etc) and also simple causal phenomena (actions, events, fluents). The Cognitive Event Calculus (CEC) (see Arkoudas’ [1]) is one such calculus which meets the above two requirements. De se beliefs are beliefs an agent has about himself/herself. Representing such beliefs is in general hard (see Rapaport’s excellent discussion in [16]), and the CEC in its present form cannot handle de se beliefs. We present simple a modification of the CEC that can let us represent de se beliefs easily.

The plan for the paper is as follows. First, we discuss the mirror test and some prior work. Then we discuss the CEC and our modifications to the calculus to represent de se beliefs. We then discuss how various types of beliefs and knowledge: de dicto, de re and de se, can be expressed in the CEC. We describe a simple architecture for agents based on the CEC. We conclude by discussing a partial formalization of the mirror test, and how it could be passed in the future by our agent named Cogito. We conclude with our intended next steps in this project.

2. THE MIRROR TEST
The sequence of a simple version of a mirror test to test for self consciousness in an animal a is as follows.

1. Anesthetize a and place a colored splotch r on a’s head
2. Awaken a and place it in front of a mirror.
3. If a removes the splotch, then declare that a is self-conscious.

Usually, a particular animal does not behave in the same manner when placed in front of a mirror. Therefore, the tests are randomized and repeated multiple times; for more details on the test see Keenan’s [8], and for a summary of relatively recent results see Prior’s work on mirror testing of magpies in [14].

The test itself is not very robust: there are many ways in the which the test can be invalid in the form of false positives and false negatives. A creature which does not have self-consciousness can pass the test in some of the following ways: 1) The splotch can be an irritant which causes the animal to scratch it. 2) The animal could be pre-trained to remove splotches on its head when placed in front of a mirror etc. Also, a self-conscious creature can fail the test.
in many ways: 1) It lacks eyesight or high level visual processing faculties. 2) It likes the splotch and prefers having it. 3) It cannot move its arms, etc.

We will not meta analyze the test itself any further; we assume prima facie the test is valid and seek to analyze what might be the capabilities required of agents that have to pass the test genuinely.

Related prior work in simulating the sensory and visual processing part of mirror image recognition can be found in Takeno’s work [17]. In this experiment, a small robot R which moves back and forth is queried to check if it discriminates among 1) R’s own mirror image; 2) another robot which imitates R; 3) another robot controlled by R; and 4) a robot which behaves randomly. This work provides evidence that at least the robotics side of the act of a simple agent recognizing itself in a mirror is feasible.

Takeno’s work deals only with the image processing part of mirror recognition and did not place or remove splotches on the robot. While the splotch removal aspect of the test is behaviourally simple, the cognitive processes involved in the test for the splotch detection and removal behaviour are not as clear as they might seem; we show this in our discussion at the end. Though the scope of our project includes a replication of the full mirror test including the visual processing part in a semi virtual character, our scholarly contribution lies in the modelling of the agent’s cognitive states as it goes through the mirror test.

3. CEC-BASED AGENTS

Our agents are based on the CEC: a multi-sorted intensional logic based on the Event Calculus. The CEC is versatile enough that it has been used to provide a formal account of mendacity in agents from beliefs about an agent who happens to be a, we need a way of distinguishing agents as actors (denoted by the res of an agent) from agents having roles (denoted by guises of an agent). This is easily understood by the analogy of a play in which different actors might have different roles on different days. The roles may change (varying guises) but the actors remain the same (constant res). That is, each agent has one and only one res but can have many guises. In CEC, the guises of agents are specified by the sort Agent and res is specified by the new sort Self and the function symbol s. The res of an agent is specified using the * function, * : Agent → Self, expressed in the postfix form as agent*, assuming that the agent expression does not contain any *. The following axioms enforce the res-guise distinction.

\[ \forall a : \text{Agent}, \exists s : \text{Self.} \ (a^s = s) \]
\[ \forall a : \text{Agent}, s_1 : \text{Self}, s_2 : \text{Self.} \ (a^s = s_1 \land a^s = s_2 \rightarrow s_1 = s_2) \]

For discussions on de se beliefs see Castañeda’s [5] and Rapaport’s [16].

4. DE DICTO, DE RE, AND DE SE

Consider our testing situation in which our agent Cogito looks at a mirror and sees a red splotch on his forehead. There are three possible ways in which we could represent Cogito’s belief when he sees the red splotch: de dicto, de re and de se. Of these, only the last one accurately captures the situation at hand.

De dicto: Cogito believes that the agent named “Cogito” has a red splotch on his head. This is represented in the CEC as follows assuming that the signature of named is named : Self × Object → Boolean.

\[ B(cogito, \exists x : \text{Agent(named(x, “Cogito”) \land red-splotched(x))}) \]

The above representation dictates that Cogito be aware of the name “Cogito.” This representation fails to differentiate our intended situation from another situation in which there is another agent named “Cogito” who has a red splotch on his head and our Cogito knows the other agent by name, and our Cogito has the above thought when seeing the other Cogito with a red splotch on his head.

De re: Cogito believes of the agent named “Cogito” that the latter has a red splotch on his head. This is represented in the CEC as follows

\[ \exists x : \text{Agent(named(x, “Cogito”) \land B(cogito, red-splotched(x))}) \]

Since we use a multi-sorted language, our quantifier variables are sorted. This is indicated as \( \forall v : \text{NameSort} \). Names are represented in the Object sort in the CEC.
This representation does not dictate that Cogito be aware of the name “Cogito.” This representation fails to differentiate our intended situation from another situation in which there is another agent named “Cogito” who has a red splotch on his head, and our Cogito has the above thought when seeing the other Cogito with a red splotch on his head.

\[ \text{De se:} \text{ Cogito believes that he himself has a red splotch on his head.} \]

This is represented in the CEC as follows

\[ B(\text{cogito, red-splotched(cogito)}) \]

Since each agent can be mapped to one and only one self symbol, we can accurately represent the situation at hand using the above simple representation.

5. AGENT ARCHITECTURE AS

The CEC gives us a way of denoting contents in agents’ minds. We still need an operational way of specifying how the agent evolves through time and interacts with its environment. Studies of human consciousness show that even though an agent may have quite a large amount of information in its mind in the form of beliefs, desires, intentions, knowledge, percepts etc. only a small subset of this information is used in acting and transitioning to the next mental state. This behaviour is usually termed executive behaviour, see Arrabales’ [2]. This is important as any non trivial agent knows a large number of facts which will not be relevant to a given situation. A simple scheme for implementing executive behavior is given now. These are extra logical notions which are used by us in implementing agents in the CEC. The notion of an agent using CEC formulae is different but closely related to the sort Agent found in the CEC.

Definition (CEC Dual Sequence) A CEC dual sequence \( A \) is any function from the natural numbers \( \mathbb{N} \) to the set of pairs of subsets of CEC expressions or formulae

\[ A : \mathbb{N} \rightarrow 2^{L_{CEC}} \times 2^{L_{CEC}} \]

\( L_{CEC} \) represents the language of the CEC, that is, the set of all formulae in the CEC.

The operational timeline is represented by the natural numbers. The function \( A \) at any moment \( t \) gives us two sets of formulae. The first element in the pair denotes the agent’s mental contents. The second element in the pair denotes the portion of the mental contents of the agent under executive control. We need a few more constraints on \( A \) to make it coherent as a mental agent.

Definition (CEC Agent Scheme AS) A CEC agent is a CEC dual sequence which satisfies these additional properties. If at any time \( t \), \( A(t) = (M_t, E_t) \) then

1. \( E_t \subseteq M_t \). The executive portion of the agents mental state is a part of the agent’s overall mental state.

2. \( M_t = (M_{t-1} - E_{t-1}) \cup E_t \). The next mental state is formed by replacing the old executive content with new content.

3. \( E_t \subseteq \{ P(a,t,\emptyset) \} \cup I(E_{t-1}) \cup M_{t-1}, \) where \( I(E) \) is the deductive closure of \( E \), that is if \( \phi \in I(E) \) then \( E \vdash \phi \). That is, the new executive content is either new perceptual information or formulae from the deductive closure of the old executive content or content from the rest of the mental space.

4. If at any time \( t \), \( I(a,t,\text{happens}(action(a,\emptyset),t)) \in E_t \) then \( \text{happens}(action(a,\emptyset),t) \) holds. That is, if the agents intends to perform an action of type \( \emptyset \), then an action of that type happens.

6. PROBLEM DISCUSSION

The question that we seek to ask is: Given that the agent sees a red splotch in the mirror on his reflection, what is that is cognitively necessary for the agent to believe that the agent himself has a red splotch on his head and then remove the splotch? We can assume that the visual processing of face detection and splotch detection on the face is carried out by lower level processing systems,
for e.g., we are using OpenCV [3] for face detection and splotch detection on faces.

That is given

\[ P(\text{cogito}, t_0, \exists x: \text{Agent. red-splotched}(x)) \in E_{t_0} \]

at time \( t_0 \) what are the other axioms necessary to have

\[ B(\text{cogito}, t_1, \text{red-splotched}(\text{cogito}*)) \in E_{t_1} \]

at time \( t_1 \) with \( t_1 \geq t_0 \). We consider two possible situations

6.1 Familiarity of Self Image

If we assume that the agent under consideration can recognize his own self image in the mirror, then the problem is trivially solved. That is if we have

\[ P(\text{cogito}, t_0, \text{red-splotched}(\text{cogito}*)) \]

We can then, trivially, get

\[ K(\text{cogito}, t_0, \text{red-splotched}(\text{cogito}*)) \]

6.2 Unfamiliarity of Self Image

The problem gets much more interesting if we remove the familiarity assumption. This can be the case for non humans, artificial agents and children who are just beginning to recognize their self images. Such agents usually experiment in front of the mirror by performing actions to see whether or not the image does the same. This goes on for some time and then there is an act of recognition based on the image repeating the agent’s actions, see [12] for experiments and discussions on these issues. We consider the case of an agent that has to learn through imitation that an agent image it is perceiving is the agent itself.

Coarsely, we could say that if agent \( a \) knows that some other agent \( a' \) performs the same actions as this agent, then both the agents are the same. This can be formalized as the Imitation Axiom A:

\[ \text{Imitation Axiom A} : \gamma_A(a) \]

\[ K(a, t, \forall \alpha: \text{ActionType}, t': \text{Moment}, \]

\[ \text{happens}(\text{action}(a, \alpha, t)) \land \text{happens}(\text{action}(a', \alpha, t')) \]

\[ \implies K(a, t, \text{mirrored}(a*) = a) \]

The problem with the Imitation Axiom A is this: how might an agent end up with the universally quantified knowledge in the antecedent? Another problem is that the agent at time \( t \) seems to know what happens in the future times \( > t \). To address these one might relax the universal quantification as follows

\[ \text{Imitation Axiom B} : \gamma_B(a) \]

\[ K(a, t, \text{happens}(\text{action}(a, \alpha, t_0)) \land \text{happens}(\text{action}(a', \alpha, t_0)) \ldots \]

\[ \land K(a, t, \text{happens}(\text{action}(a, \alpha, t_0 + n)) \land \text{happens}(\text{action}(a', \alpha, t_0 + n)) \]

\[ \land t \geq t_0 + n \implies K(a, t, \text{mirrored}(a*) = a') \]

6.3 Next Steps: Cogito’s Mental Contents

Our next step involves specifying the contents of \( E_{t_0} \) for our agent system and show that there exists a \( t_f \) such that

\[ I(\text{cogito}, t_f, \text{happens}(\text{action}(\text{cogito*}, \text{remove-splotch}), t_f)) \in E_{t_f} \]

At the minimum we need to assume that Cogito has some knowledge of the external world, that is knowledge concerning some physics of the world formalized via the Event Calculus. For this purpose, we introduce the following function symbols

\[ \text{has-splotch} : \text{Self} \rightarrow \text{ Fluent} \]

\[ \text{remove-splotch} : \text{Self} \rightarrow \text{ActionType} \]

Very trivially, Cogito should know that the action of removing the splotch from and by himself leads to the fluent \( \text{has-splotch} \) not holding.

\[ \phi_1 \equiv \forall t : \text{Moment. } \left( K(t, \text{removes}(\text{action}(\text{cogito*}, \text{remove-splotch}(\text{cogito*})), t) \right) \]

Also for the action part of Cogito to work, we need that Cogito intends to not have the splotch on his forehead at all times.

\[ \phi_2 \equiv \forall t : \text{Moment. } I(t) \left( \text{cogito}, t, \forall t : \text{Moment. } \neg \text{holds}(\text{has-splotch}(\text{cogito*}), t) \right) \]

We need to account for the agent’s knowledge that properties that hold for its mirror image hold for itself too. We have the following axiom schema:

\[ \mu(a, F) \equiv \forall t, t': \text{Moment}, \forall a : \text{Agent} \left( B(a, t', \text{mirrored}(a*) = a') \implies \right. \]

\[ \left. B(t, \text{holds}(F(a*), t)) \leftrightarrow \text{holds}(F(a'), t) \right) \]

At the least, now one can conclude that:

\[ \{ \gamma_B(\text{cogito*}), \mu(\text{cogito, has-splotch}), \phi_1, \phi_2 \} \subseteq E_{t_0} \]

The rest of \( E_{t_0} \) needs to be determined and is the subject of ongoing research.

7. REFERENCES


