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 ($\mathcal{CEC}$). We present a simple agent specification formalism, $\mathcal{AS}$, based on the $\mathcal{CEC}$. 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 ($\mathcal{CEC}$) to handle de se beliefs in a fashion paralleling Castañeda's suggestions. The main contributions of this paper are 1) $\mathcal{AS}$: a simple and formal specification of a $\mathcal{CEC}$-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 here[1]. 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 ($\mathcal{CEC}$) (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 $\mathcal{CEC}$ in its present form cannot handle de se beliefs. We present simple a modification of the $\mathcal{CEC}$ 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 $\mathcal{CEC}$ 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 $\mathcal{CEC}$. We describe a simple architecture for agents based on the $\mathcal{CEC}$. 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

[1] We do not claim that our agent can be capable of any form of consciousness. This is a philosophically rich and controversial field which we set aside for philosophers.

Cite as: Towards a Logic-based Analysis and Simulation of the Mirror Test, 5,

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 on mirror image discrimination in [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 $\mathcal{CEC}$: a multi-sorted intensional logic based on the Event Calculus. The $\mathcal{CEC}$ is versatile enough that it has been used to provide a formal account of mendacity in agents by Clark in [6] and to analyze the false-belief task by Arkoudas and Bringsjord in [1]. We give a quick overview of the $\mathcal{CEC}$; for a more detailed explanation please see [1] or Bringsjord's [4]. The previous tasks are similar in nature to the mirror recognition task and this similarity is one of the reasons why we chose the $\mathcal{CEC}$ over similar systems described in [7, 15, 10, 9].

3.1 The Cognitive Event Calculus

In this paper we use Version 2 of the $\mathcal{CEC}$. The syntax of Version 2 of the $\mathcal{CEC}$ and some of its inference rules are shown in Figure 1. The CEC is a multi-sorted first-order modal logic in which modal operators specify intentional contexts. The operators are: the knowledge operator **K**, the belief operator **B**, the percept operator **P**, the intention operator **I**, the desire operator **D**, the communication operator **S** and the common knowledge operator **C**. Version 2 of $\mathcal{CEC}$ differs from the Version 1 in 1) having time indexed modal operators; 2) the operators **D**, **I** and **S**; and 3) machinery for de se beliefs. Only the last addition concerns our purpose here.[2] We now give a brief informal interpretation of the calculus.

We denote that agent a knows ϕ at time t by $\mathbf{K}(a,t,\phi)$. The operators **B**, **P**, **D** and **I** have a similar informal interpretation. The formula $\mathbf{S}(a,b,t,\phi)$ captures declarative communication of ϕ from agent a to agent b at time t. Common-knowledge of ϕ in the system is denoted by $C(t,\phi)$. Common-knowledge of some proposition ϕ holds when every agent knows ϕ and every agent knows that every agent knows ϕ and so on ad infinitum. The $\textsf{Moment}$ sort is used for representing time points. We assume that the time points are isomorphic with $\mathbb{N}$ and functions $+, -$, relations $>, <, \geq, \leq$ are available.

The $\mathcal{CEC}$ includes the signature of the classic Event Calculus (EC), see Mueller's [11], and the axioms of EC are assumed to be common knowledge in the system [1]. The EC is a first-order calculus that lets us reason about events that occur in time and their effects on fluents.

3.2 De se beliefs: res and guise

In order to represent and distinguish self beliefs by an agent a 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 $\mathcal{CEC}$, the guises of agents are specified by the sort $\textsf{Agent}$ and res is specified by the new sort $\textsf{Self}$ and the function symbol $*$. The res of an agent is specified using the $*$ function, $* : \textsf{Agent} \rightarrow \textsf{Self}$, expressed in the postfix form as $agent*$, assuming that the *agent* expression does not contain any $*$. The following axioms enforce the res-guise distinction.

$$\forall a : \textsf{Agent}, \exists s : \textsf{Self}.\ (a* = s)$$

$$\forall a : \textsf{Agent}, s_1 : \textsf{Self}, s_2 : \textsf{Self}.(a* = s_1 \land a* = s_2 \rightarrow s_1 = s_2)$$

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 $\mathcal{CEC}$ as follows assuming that the signature of *named* is *named* : $\textsf{Self} \times \textsf{Object} \rightarrow \textsf{Boolean}$. [3]

$$\mathbf{B}(cogito, \exists x : \textsf{Agent}(named(x, \text{``}Cogito\text{''}) \land red\text{-}splotched(x)))$$

The above representation dictates that Cogito be aware of the name "Cogito."[4] 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: *Cogtio believes of the agent named "Cogito" that the latter has a red splotch on his head.* This is represented in the $\mathcal{CEC}$ as follows

$$\exists x : \textsf{Agent}(named(x, \text{``}Cogito\text{''}) \land \mathbf{B}(cogito, red\text{-}splotched(x)))$$

[2] We refrain from specifying a formal semantics for the calculus as we feel that the possible worlds approach, the only candidate which is precise enough, falls short of the *tripartite analysis of knowledge* (Pappas [13]). In the tripartite analysis, knowledge is a belief which is true and justified. The standard possible worlds semantics for epistemic logics skips over the justification criterion for knowledge. Instead of a formal semantics for our calculus, we specify a set of inference rules that capture our informal understanding and semantics underlying the calculus.

[3] Since we use a multi-sorted language, our quantifier variables are sorted. This is indicated as $\exists var : \textsf{SortName}$

[4] Names are represented in the $\textsf{Object}$ sort in the $\mathcal{CEC}$.

Figure 1: Cognitive Event Calculus (Version 2)

Syntax

$$
\begin{aligned}
S ::= {} & \textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \\
& \textsf{Moment} \mid \textsf{Boolean} \mid \textsf{Fluent} \mid \textsf{RealTerm} \\
& \mathit{action} : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action} \\
& \mathit{initially} : \textsf{Fluent} \rightarrow \textsf{Boolean} \\
& \mathit{holds} : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
& \mathit{happens} : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
f ::= {} & \mathit{clipped} : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \mathit{Boolean} \\
& \mathit{initiates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
& \mathit{terminates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
& \mathit{prior} : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\
& \mathit{interval} : \textsf{Moment} \times \textsf{Boolean} \\
& * : \textsf{Agent} \rightarrow \textsf{Self} \\
t ::= {} & x : S \mid c : S \mid f(t_1, \ldots, t_n) \\
& t : \textsf{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \\
\phi ::= {} & \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \\
& \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\phi) \mid \mathbf{I}(a,t,\phi) \mid \mathbf{S}(a,b,t,\phi)
\end{aligned}
$$

Rules of Inference

$$\frac{}{\mathbf{C}(t, \mathbf{P}(a,t,\phi) \Rightarrow \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t, \mathbf{K}(a,t,\phi) \Rightarrow \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq l_n}{\mathbf{K}(a_1, t_1, \ldots k(a_n, t_n, \phi) \ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{}{\mathbf{C}(\mathbf{C}(t, \mathbf{K}(a, t_1, \phi_1 \Rightarrow \phi_2)) \Rightarrow \mathbf{K}(a, t_2, \phi_1) \Rightarrow \mathbf{K}(a, t_3, \phi_3))}\ [R_5]$$

$$\frac{}{\mathbf{C}(\mathbf{C}(t, \mathbf{B}(a, t_1, \phi_1 \Rightarrow \phi_2)) \Rightarrow \mathbf{B}(a, t_2, \phi_1) \Rightarrow \mathbf{B}(a, t_3, \phi_3))}\ [R_6]$$

$$\frac{}{\mathbf{C}(\mathbf{C}(t, \mathbf{C}(t_1, \phi_1 \Rightarrow \phi_2)) \Rightarrow \mathbf{C}(t_2, \phi_1) \Rightarrow \mathbf{C}(t_3, \phi_3))}\ [R_7]$$

$$\frac{}{\mathbf{C}(t, \forall x.\ \phi \Rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t, \phi_1 \Leftrightarrow \phi_2 \Rightarrow \neg\phi_2 \Rightarrow \neg\phi_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}(t, [\phi_1 \wedge \ldots \wedge \phi_n \Rightarrow \phi] \Rightarrow [\phi_1 \Rightarrow \ldots \Rightarrow \phi_n \Rightarrow \phi])}\ [R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi_1)\ \mathbf{B}(a,t,\phi_2)}{\mathbf{B}(a,t,\phi_1 \wedge \phi_2)}\ [R_{11}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a,\alpha),t))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a,\alpha),t))}\ [R_{13}]$$

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.

De se: *Cogtio believes that he himself has a red splotch on his head*. This is represented in the $\mathcal{CEC}$ as follows

$$\mathbf{B}(\mathit{cogito}, \mathit{red\text{-}splotched}(\mathit{cogito}*)))$$

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 $\mathcal{AS}$

The $\mathcal{CEC}$ 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 $\mathcal{CEC}$. The notion of an agent using $\mathcal{CEC}$ formulae is different but closely related to the sort Agent found in the $\mathcal{CEC}$.

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

$$\mathbf{A} : \mathbb{N} \mapsto 2^{\mathcal{L}_{\mathcal{CEC}}} \times 2^{\mathcal{L}_{\mathcal{CEC}}}$$

$\mathcal{L}_{\mathcal{CEC}}$ represents the language of the $\mathcal{CEC}$, that is, the set of all formulae in the $\mathcal{CEC}$.

The operational timeline is represented by the natural numbers. The function $\mathbf{A}$ 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 $\mathbf{A}$ to make it coherent as a mental agent.

Definition (CEC Agent Scheme $\mathcal{AS}$) *A $\mathcal{CEC}$ agent is a $\mathcal{CEC}$ dual sequence which satisfies these additional properties. If at any time t, $\mathbf{A}(t) = \langle \mathbf{M}_t, \mathbf{E}_t \rangle$ then*

1. $\mathbf{E}_t \subseteq \mathbf{M}_t$. *The executive portion of the agents mental state is a part of the agent's overall mental state.*

2. $\mathbf{M}_t = (\mathbf{M}_{t-1} - \mathbf{E}_{t-1}) \cup \mathbf{E}_t$. *The next mental state is formed by replacing the old executive content with new content.*

3. $\mathbf{E}_t \subseteq \{\mathbf{P}(a,t,\phi)\} \cup \mathbf{I}(\mathbf{E}_{t-1}) \cup \mathbf{M}_{t-1}$, *where* $\mathbf{I}(\mathbf{E})$ *is the deductive closure of* $\mathbf{E}$, *thats is if* $\phi \in \mathbf{I}(\mathbf{E})$ *then* $\mathbf{E} \vdash \phi$. *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,* $\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a,\alpha),t)) \in \mathbf{E}_t$ *then* $\mathit{happens}(\mathit{action}(a,\alpha),t))$ *holds. That is, if the agents intends to perform an action of type* α, *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 it 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

$$\mathbf{P}(cogito, t_0, \exists x : \mathsf{Agent}.\ red\text{-}splotched(x)) \in \mathbf{E}_{t_0}$$

at time t_0 what are the other axioms necessary to have

$$\mathbf{B}(cogito, t_1, red\text{-}splotched(cogito*)) \in \mathbf{E}_{t_1}$$

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

$$\mathbf{P}(cogito, t_0, red\text{-}splotched(cogito*))$$

We can the use $[DR_4]$ and $[DR_5]$ of the $\mathcal{CEC}$(see [1])to derive our intended belief.

$$\frac{\mathbf{P}(a,t,\phi)}{\mathbf{K}(a,t,\phi)}\ [DR_4] \qquad\qquad \frac{\mathbf{K}(a,t,\phi)}{\mathbf{B}(a,t,\phi)}\ [DR_5]$$

We can then, trivially, get

$$\cfrac{\cfrac{\mathbf{P}(cogito, t_0, red\text{-}splotched(cogito*))}{\mathbf{K}(cogito, t_0, red\text{-}splotched(cogito*))}\ [DR_4]}{\mathbf{B}(cogito, t_0, red\text{-}splotched(cogito*))}\ [DR_5]$$

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:

$$\underline{\mathsf{Imitation\ Axiom\ A}} : \gamma_A(a)$$

$$\Big[\mathbf{K}\big(a, t, \forall\alpha : \mathsf{ActionType},\ t : \mathsf{Moment},$$
$$happens(action(a,\alpha),t) \wedge happens(action(a',\alpha),t)\big)\Big]$$
$$\Longrightarrow \mathbf{K}(a,t,mirrored(a*) = a)$$

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

$$\underline{\mathsf{Imitation\ Axiom\ B}} : \gamma_B(a)$$

$$\Big[\mathbf{K}\big(a, t, happens(action(a,\alpha_1),t_0) \wedge happens(action(a',\alpha),t_0)\big)\ldots$$
$$\wedge \mathbf{K}\big(a, t, happens(action(a,\alpha_n),t_0+n) \wedge happens(action(a',\alpha_n),t_0+n)\big)$$
$$\wedge\, t \geq t_0 + n\Big] \Longrightarrow \mathbf{K}(a,t,mirrored(a*) = a')$$

6.3 Next Steps: Cogito's Mental Contents

Our next step involves specifying the contents of $\mathbf{E}_{t_0}$ for our agent system and show that there exists a $t_f > t_0$ such that

$$\mathbf{I}(cogito, t_f, happens(action(cogito*, remove\text{-}splotch), t_f)) \in \mathbf{E}_{t_f}$$

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

$$has\text{-}splotch : \mathsf{Self} \rightarrow \mathsf{Fluent}$$
$$remove\text{-}splotch : \mathsf{Self} \rightarrow \mathsf{ActionType}$$

Very trivially, Cogito should know that the action of removing the splotch from and by himself leads to the fluent *has-splotch* not holding.

$$\phi_1 \equiv \forall tt : \mathsf{Moment}.\ \mathbf{K}\left(\begin{array}{l} cogito, tt, \forall t : \mathsf{Moment}. \\ terminates(action(cogito*, \\ \qquad\qquad remove\text{-}splotch(cogito*)), \\ \qquad\qquad has\text{-}splotch(cogito*), t) \end{array}\right)$$

Also for the action part of Cogito to work, we need that Cogito intends to not have the splotch on his forehead. ϕ_2 formalizes the fact Cogito intends that the splotch not be on his forehead at all times.

$$\phi_2 \equiv \forall tt : \mathsf{Moment}.\ \mathbf{I}\left(\begin{array}{l} cogito, tt, \\ \forall t : \mathsf{Moment}. \neg holds(has\text{-}splotch(cogito*), t) \end{array}\right)$$

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:

$$\mu(a,F) \equiv \forall tt, t : \mathsf{Moment}, \forall a : \mathsf{Agent} \left(\begin{array}{c} \mathbf{B}(a,t,mirrored(a*) = a') \Longrightarrow \\ \left(\begin{array}{c} \mathbf{B}(t, holds(F(a'),tt) \leftrightarrow \\ holds(F(a*),tt)) \end{array}\right) \end{array}\right)$$

At the least, now one can conclude that:

$$\{\gamma_B(cogito), \mu(cogito, has\text{-}splotch), \phi_1, \phi_2\} \subseteq \mathbf{E}_{t_0}$$

The rest of $\mathbf{E}_{t_0}$ needs to be determined and is the subject of ongoing research.

7. REFERENCES

[1] K. Arkoudas and S. Bringsjord. Toward formalizing common-sense psychology: An analysis of the false-belief task. *PRICAI 2008: Trends in Artificial Intelligence*, pages 17–29, 2008.

[2] R. Arrabales, A. Ledezma, and A. Sanchis. Criteria for consciousness in artificial intelligent agents. In *Proceeding of: Adaptive Learning Agents and Multi-Agent Systems, ALAMAS+ALAg 2008-Workshop at AAMAS 2008, Estoril, May, 12, 2008, Portugal.*., pages 57–64, 2008.

[3] G.R. Bradski and A. Kaehler. *Learning opencv*. O'Reilly, 2008.

[4] S. Bringsjord. Meeting floridi's challenge to artificial intelligence from the knowledge-game test for self-consciousness. *Metaphilosophy*, 41(3):292–312, 2010.

[5] H.N. Castañeda. On the logic of self-knowledge. *Nous*, 1(1):9–21, 1967.

[6] M. Clark. *Cognitive Illusions and the Lying Machine*. PhD thesis, PhD thesis, Rensselaer Polytechnic Institute (RPI), 2008.

[7] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261, 1990.

[8] J.P. Keenan, G.C. Gallup, and D. Falk. *The face in the mirror: The search for the origins of consciousness.* HarperCollins Publishers, 2003.

[9] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.

[10] J.J. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence Preprint Series*, 14, 1999.

[11] E.T. Mueller. *Commonsense reasoning*. Morgan Kaufmann, 2006.

[12] M. Nielsen and C. Dissanayake. Pretend play, mirror self-recognition and imitation: A longitudinal investigation through the second year. *Infant Behavior and Development*, 27(3):342–365, 2004.

[13] G.S. Pappas and M. Swain. *Essays on knowledge and justification*. Cornell University Press, 1978.

[14] H. Prior, A. Schwarz, and O. Güntürkün. Mirror-induced behavior in the magpie (pica pica): evidence of self-recognition. *PLoS Biology*, 6(8):e202, 2008.

[15] A.S. Rao, M.P. Georgeff, Austrialian Artificial Intelligence Institute, Technology Department of Industry, and Australia Commerce. *Modeling rational agents within a BDI-architecture*. Australian Artificial Intelligence Institute, 1991.

[16] W.J. Rapaport, S.C. Shapiro, and J.M. Wiebe. Quasi-indexicals and knowledge reports. *Cognitive Science: A Multidisciplinary Journal*, 21(1):63–107, 1997.

[17] J. Takeno, K. Inaba, and T. Suzuki. Experiments and examination of mirror image cognition using a small robot. In *Computational Intelligence in Robotics and Automation, 2005. CIRA 2005. Proceedings. 2005 IEEE International Symposium on*, pages 493–498. IEEE, 2005.