Real Robots that Pass Human Tests of Self-Consciousness

Selmer Bringsjord • John Licato • Naveen Sundar Govindarajulu • Rikhiya Ghosh • Atriya Sen
Rensselaer AI & Reasoning (RAIR) Lab
Department of Computer Science • Department of Cognitive Science
Rensselaer Polytechnic Institute (RPI) • Troy NY 12180 USA
Contact: Bringsjord (selmer@rpi.edu)

***Abstract*—**
Self-consciousness would seem to be a *sine qua non* for moral competence in a social world. You and we are morally competent in no small part because you know what you ought to do, and we know what we ought to do. A mouse, in contrast, cannot say to itself: "I ought to share this cheese, even if my brother refuses to do so." But can *robots* be self-conscious? Approaching this question from the standpoint of so-called *Psychometric AI*, we note that prior work by Govindarajulu and Bringsjord led to the engineering of a robot (Cogito) able to *provably* pass the famous mirror test of self-consciousness. But a more challenging test for robot self-consciousness has been provided by Floridi; this test is an ingenious and much-harder variant of the well-known-in-AI wise-man puzzle: Each of three robots is given one pill from a group of five, three of which are innocuous, but two of which, when taken, immediately render the recipient dumb. In point of fact, two robots (R_1 and R_2) are given potent pills, but R_3 receives one of the three placebos. The human tester says: "Which pill did you receive? No answer is correct unless accompanied by a proof!" Given a formal regimentation of this test previously formulated by Bringsjord, it can be proved that, in theory, a future robot represented by R_3 can answer provably correctly (which for plausible reasons, explained by Floridi, entails that R_3 has satisfied some of the structural requirements for self-consciousness). In this paper we explain and demonstrate the engineering that now makes this theoretical possibility actual, both in the simulator known as 'PAGI World' (used for testing AIs), and in real (= physical) robots interacting with a human tester. These demonstrations involve scenarios that demand the passing of Floridi's test for self-consciousness, where for us, passing such a test is required for an agent to be regarded morally competent.

I. Introduction

Self-consciousness would seem to be a *sine qua non* for moral competence in a social world. You and we are morally competent in no small part because you know what you ought to do, and we know what we ought to do. A mouse, in contrast, cannot say to itself: "I ought to share this cheese, even if my brother refuses to do so." Or to consider a more relevant case: If Black threatens to shoot you if you don't go into a nearby store and shoplift a candy bar for him, it wouldn't really be *you* who steals the candy bar; rather, Black would be the blameworthy one; and this diagnosis presupposes self-consciousness, at least in some form. In addition, moral competence in a robot situated among humans clearly requires sophisticated and natural human-robot interaction, of the sort envisioned by Scheutz [1], and such interaction will require that the robot be able to (among other things) discuss, in natural language, self-ascriptions and self-control in connection with morality. For example, blame, under investigation by Malle [2], is a key concept in human moral discourse, and obviously such claims as "I am not to blame" are bound up inextricably with at least *structures* relating to self-consciousness.[1]

But can robots *be* self-conscious? From the standpoint of *Psychometric AI* [4], [5], [6], [7], which, consistent with the spirit of the Turing Test [8], reduces such deeply puzzling and controversial philosophical questions as this one to concrete engineering effort focused on building agents/robots that can pass well-defined tests, this question becomes: Can robots pass test $\mathcal{T}_{\text{S-C}}$ for self-consciousness? Prior Psychometric-AI work on this question by Govindarajulu and Bringsjord [9], [10] led to the engineering of a robot (Cogito) able to *provably* pass the famous mirror test of self-consciousness. But a much more challenging test for robot self-consciousness has been provided by Floridi [11]; this test is an ingenious and much-harder variant of the well-known-in-AI wise-man puzzle [which is discussed along with other such cognitize puzzles e.g. in [12]]: Each of three robots is given one pill from a group of five, three of which are innocuous, but two of which, when taken, immediately render the recipient dumb. In point of fact, two robots (R_1 and R_2) are given potent pills, but R_3 receives one of the three placebos. The human tester says: "Which pill did you receive? No answer is correct unless accompanied by a proof!" Given a formal regimentation of this test formulated and previously published by Bringsjord [13], it can be proved that, in theory, a future robot represented by R_3 can answer provably correctly (which for reasons given

We thank Luciano Floridi, as without his seminal reflection and writing on robots and self-consciousness, the trajectory of AI r&d on which we report herein would never have been born. With deep gratitude, we acknowledge that our work is supported by an ONR MURI originally sagaciously overseen on the DoD side by P Bello, and now in like manner by Micah Clark. Some support as well from AFOSR and IBM has also been very helpful, and we are thankful for this too. In addition, without the support of, and interaction with, our energetic and brilliant co-investigators in ONR-sponsored work (i.e. Co-PIs B Malle and M Sei, and PI M Scheutz (MURI), and PI R Sun (moral reasoning)), what we present herein would be — to severely understate — severely compromised, and so on this front too we express profound gratitude. Finally, we are grateful to two anonymous reviewers for a number of trenchant comments and objections.

[1] On the rationale for the mere focus on the structural aspects of self-consciousness, see §II. For excellent work that is at once structural/computational, and, unlike that displayed in the present paper, informed by cognitive neuro/science, see [3].

by Floridi entails that R_3 has confirmed structural aspects of self-consciousness). In the present paper we explain and demonstrate the engineering that now makes this theoretical possibility actual, both in the simulator known as 'PAGI World' (used for testing AIs), and in real (= physical) robots interacting with a human tester. These demonstrations involve scenarios that demand, from agents who would pass, behavior that suggests that self-consciousness in service of morally competent decision-making is present.

The present paper's plan: We begin (in §II) with a deflationary disclaimer in which we explain that we are doing engineering, not philosophy. Next, in section III, we very briefly recount work on the mirror test. Then (§V) we describe the promised PAGI-World demonstration. After that, in section VI, we move from simulation to physical robots, and show that Floridi's test can be met in real time by sufficiently "self-conscious" robots. We draw the paper to close (§VIII) by announcing the next steps in our research program, intended to be taken by the time RO-MAN 2015 occurs.

II. Disclaimer: Tests and Structure Only

Bringsjord doesn't believe that any of the artificial creatures featured in the present paper are *actually* self-conscious. He has explained repeatedly [e.g., see [14], [15]] that genuine *phenomenal* consciousness [16] is impossible for a mere machine to have, and true self-consciousness would require phenomenal consciousness. *Nonetheless, the logico-mathematical structure and form of self-consciouness can be ascertained and specified, and these specifications can then be processed computationally in such a way as to meet clear tests of mental ability and skill.* This test-based approach, dubbed Psychometric AI, thankfully, avoids endless philosophizing in favor of determinate engineering aimed at building AIs that can pass determinate tests. In short, computing machines, AIs, robots, and so on are all "zombies," but these zombies can be engineered to pass tests. A not-small body of work lays out and establishes this position; e.g., [14], [17], [18], [19], [5], [4]. Some of Bringsjord's co-authors in the present case may well reject his position, but no matter: engineering to tests is fortunately engineering, not a matter of metaphysics.

III. Mirror-Test Engineering

Figure 1 shows the set of axioms Γ_1 that were used in a simulation of Cogito, in which passing of the test is secured. Some $\mathcal{DCEC}^*$ formulae (not shown here) also connect knowledge, belief, desire, perception, and communication. For a full discussion, see [20]. At RO-MAN 2015, demonstration of success on the mirror test will be provided.

IV. $\mathcal{DCEC}^*$

The Deontic Cognitive Event Calculus ($\mathcal{DCEC}^*$) is a logicist framework supporting a multi-sorted, quantified modal logic [21]. $\mathcal{DCEC}^*$ contains operators for belief, knowledge, intention, obligation, and others, thus allowing the representation of doxastic (belief) and deontic (obligation) formulae.

Fig. 1: Propositions Used in Mirror-Test Simulation

Imitation If I see another agent a perform the same actions as me twice concurrently, then I know that the other agent is my mirror reflection

$$\begin{aligned}&\texttt{Imit}: \forall(t_1, t_2 : \text{Moment}, a : \text{Agent}, act_1, act_2 : \text{Action})\\ &\Big(\mathbf{K}(\mathsf{I}, t_1, happens(action(\mathsf{I}, act_1), t_1)) \wedge \mathbf{K}(\mathsf{I}, t_1, happens(action(a, act_1), t_1))\\ &\mathbf{K}(\mathsf{I}, t_2, happens(action(\mathsf{I}, act_2), t_2)) \wedge \mathbf{K}(\mathsf{I}, t_2, happens(action(a, act_2), t_2))\Big)\\ &\rightarrow \mathbf{K}(\mathsf{I}, \text{now}, mirror(\mathsf{I}, a))\end{aligned}$$

Wave Left I know that I wave left at time t_1 and I can perceive this action of mine.

$$\begin{aligned}\texttt{Wave}_{\texttt{left}}: &\mathbf{K}(\mathsf{I}, t_1, happens(action(\mathsf{I}, wave_{left}), t_1)) \wedge\\ &\mathbf{P}(\mathsf{I}, t_1, happens(action(\mathsf{I}, wave_{left}), t_1))\end{aligned}$$

Wave Right I know that I wave right at time t_2 and I can perceive this action of mine.

$$\begin{aligned}\texttt{Wave}_{\texttt{right}}: &\mathbf{K}(\mathsf{I}, t_2, happens(action(\mathsf{I}, wave_{right}), t_2)) \wedge\\ &\mathbf{P}(\mathsf{I}, t_2, happens(action(\mathsf{I}, wave_{right}), t_2))\end{aligned}$$

Mirror Physics If I see another agent a with a red splotch on its head, and if I believe that the other agent is my mirror reflection, then I believe that I too have a red splotch.

$$\begin{aligned}&\texttt{Physics}_{\texttt{mirror}}: \forall(a : \text{Agent})\\ &(\mathbf{P}(\mathsf{I}, \text{now}, holds(red\text{-}splotched(a), \text{now})) \wedge \mathbf{B}(\mathsf{I}, \text{now}, mirror(\mathsf{I}, a)))\\ &\rightarrow \mathbf{B}(\mathsf{I}, \text{now}, holds(red\text{-}splotched(\mathsf{I}*), \text{now}))\end{aligned}$$

Wipe Action I know that if I myself wipe my own forehad, the splotch will be gone .

$$\texttt{Wipe}_{\texttt{action}}: \mathbf{K}(\mathsf{I}, \text{now}, terminates(action(\mathsf{I}*, wipe\text{-}fore\text{-}head(\mathsf{I}*)), red\text{-}splotched(\mathsf{I}*), \text{now}))$$

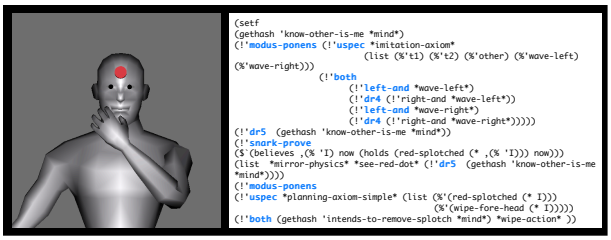
Planning A simple planning axiom.

$$\begin{aligned}&\texttt{Planning}: \forall(f : \text{Fluent}, \alpha : \text{ActionType})\\ &\mathbf{I}(\mathsf{I}, \text{now}, \neg holds(f, \text{now})) \wedge \mathbf{K}(\mathsf{I}, \text{now}, terminates(action(\mathsf{I}*, \alpha), f, \text{now}))\\ &\rightarrow \mathbf{I}(\mathsf{I}, \text{now}, happens(action(\mathsf{I}*, \alpha), \text{now}))\end{aligned}$$

No Splotch I do not want the splotch.

$$\begin{aligned}\texttt{No}_{\texttt{splotch}}: &\forall(t : \text{Moment})\mathbf{D}(\mathsf{I}, t, \neg holds(red\text{-}splotched(\mathsf{I}*), t)) \wedge\\ &\mathbf{B}(\mathsf{I}, t, \neg holds(red\text{-}splotched(\mathsf{I}*), t))\end{aligned}$$

Fig. 2: Cogito Removing the Dot, a Part of the Simulation



Recently, RAIR Lab researchers have been developing an automated theorem prover for $\mathcal{DCEC}^*$, an early version of which is used in Section V. The current syntax and rules of inference for the simple dialect of $\mathcal{DCEC}^*$ used herein are shown in Figure 3.

$\mathcal{DCEC}^*$ differs from Belief-Desire-Intention (BDI) logics [22] in many important ways (see [23] for a discussion). For example, $\mathcal{DCEC}^*$ explicitly rejects possible-worlds semantics and model-based reasoning, instead opting for a *proof-theoretic* semantics and the associated type of reasoning commonly referred to as *natural deduction* [24], [25], [26], [27]. In addition, as far as we know, $\mathcal{DCEC}^*$ is the only family of logics in which desiderata regarding the personal pronoun I^* laid down by deep theories of self-consciousness [e.g., see [28]], are provable theorems. For instance it is a theorem that if some agent a has a first-person belief that I_a^* has some attribute R, then no formula expressing that some term t has R can be proved. This a requirement because, as [28] explains, the distinctive nature of first-person

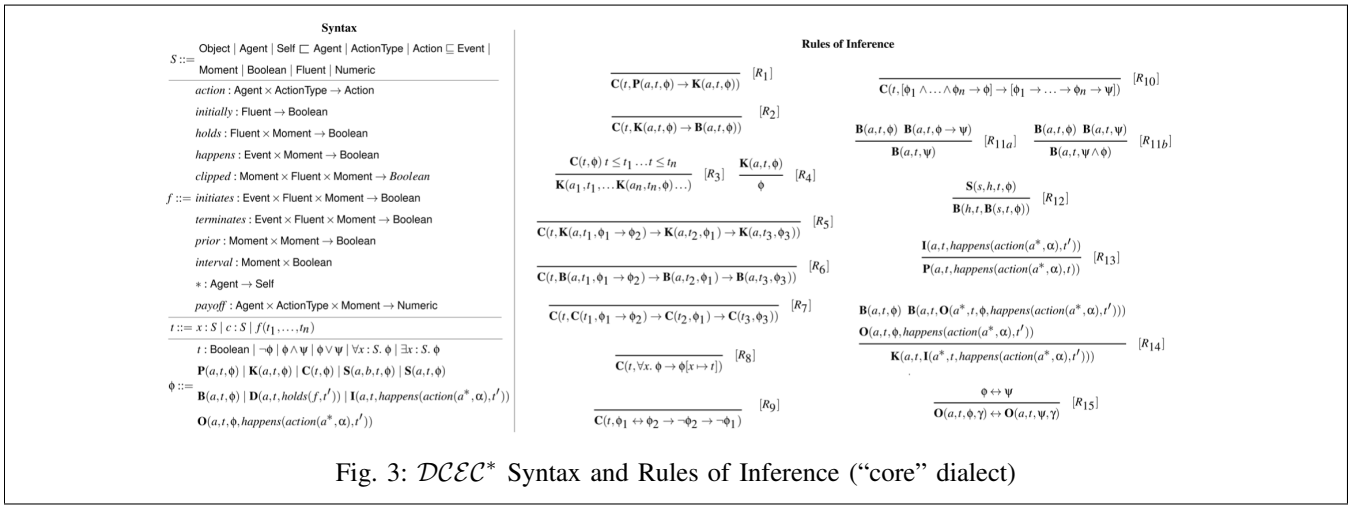


Fig. 3: $\mathcal{DCEC}^*$ Syntax and Rules of Inference ("core" dialect)

consciousness is that one can have beliefs about oneself in the complete absence of bodily sensations.

V. Demonstration in PAGI World

In order to show the initial demonstration, we made use of PAGI (pronounced "pay-guy") World, a simulation environment developed by the RAIR Lab for the testing and development of artificially intelligent agents. PAGI World is built out of the game-development engine Unity3d, and is designed to be extremely easy for AI researchers to work with. It achieves its ease-of-use by being open-sourced, able to run on all major platforms (Windows, MacOS, and most Linux distributions), free to use, and able to be controlled by almost any programming language. Since PAGI World communicates with AI controllers through TCP/IP, theoretically any language which can send strings over TCP/IP can serve as AI controllers, interacting with PAGI World by sending and receiving low-level information. For example, the AI controller can send commands to send downward force to the hands of the AI agent in the PAGI World environment (whom we usually refer to as the "PAGI Guy"). If one of the hands touches an object in the environment, sensory data will be sent back from PAGI World to the AI controller (through TCP/IP) containing basic information like the approximate temperature of the object, which sensor on the hand was hit by the object, and so on. Figure 4 shows the overall architecture of PAGI World and a typical AI controller (which we sometimes refer to as the 'PAGI-side' and the 'AI-side,' respectively).

Since PAGI World draws on Unity3d's physics engine, PAGI World tasks can incorporate realistic physics (though only a 2-dimensional physics is used for simplicity). A text box is optionally provided in PAGI World, so that a human controller can type text in PAGI World which will be sent to the AI-side and processed as if it were a statement uttered to PAGI Guy. A text display in PAGI World can also display messages sent from the AI-side to PAGI World, to emulate PAGI Guy "speaking." In the AI controller pictured in Figure 4, text sent to and from the AI-side can be parsed to, and converted from, formulae in $\mathcal{DCEC}^*$.

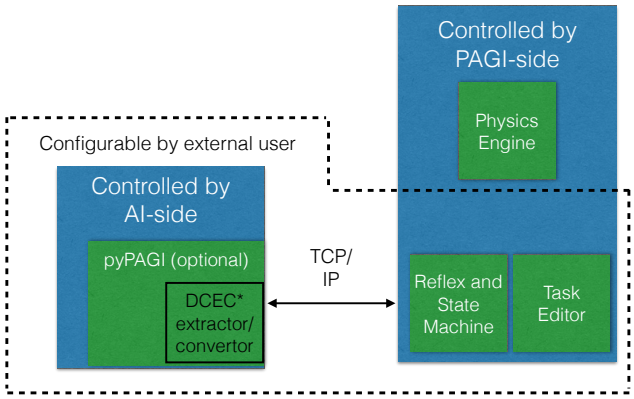


Fig. 4: The Architecture of PAGI World (the "PAGI-side") and Typical AI Controller (the "AI-side"). Note that the details of the AI-side are completely up to the AI programmer.

A. Floridi's KG4 (= Dumbing Pill Test) in PAGI World

We can now describe the task that simulates success in Floridi's self-consciousness test. Following [29], we create a task in which three robots, one of them PAGI Guy, are in a room with five pills (Figure 5). Three of these pills are mere placebos, but the other two are "dumbing" pills, meaning they make the robot who ingests them unable to speak. The pills are visually distinguishable to a human observer — the dumbing pills are colored red — but this information is not accessible to the robots.

Prior to the start of the task (at time t_1 = *"apprise"*), the robots are given knowledge about how the task works in the form of $\mathcal{DCEC}^*$ formulae. At time t_2 = *"ingest"*, the human controller drags the pills and drops one on each robot (Figure 6), which then ingests the pill. The pills are selected randomly by the human controller, and the robots are all given knowledge that they will be given pills at t_2 (but not knowledge of which pill they will be given). At time t_3 = *"inquire"*, the human controller opens the text box in PAGI World and types in the following (without the line break):

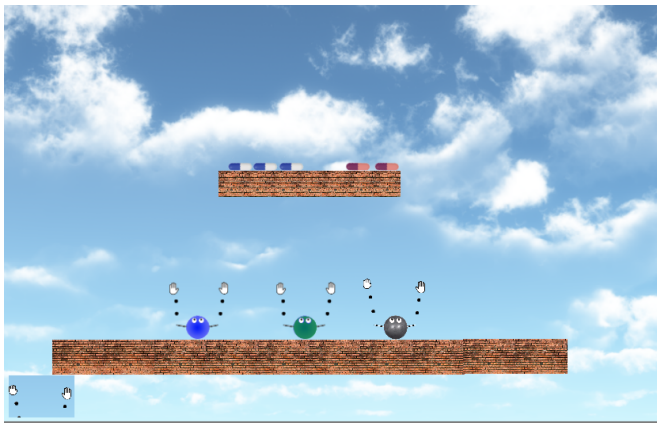
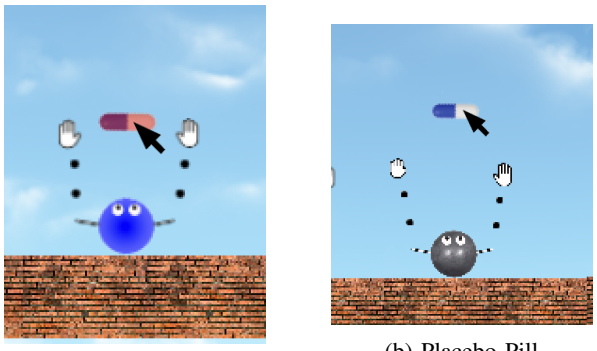


Fig. 5: The Task in PAGI World in its Starting Configuration



(a) Dumbing Pill (b) Placebo Pill

Fig. 6: The Robots Being Given Pills

```
K(R3,t4, not(happens
     (action(R3,ingestDumbPill),t2)))?
```

This text is sent to the AI controller and converted into a $\mathcal{DCEC}^*$ formula ϕ. R_3, the robot whose knowledge is being queried, is the label assigned to PAGI Guy, who in our experiment is given the placebo pill. The question-mark is interpreted as a command to attempt to answer whether or not ϕ holds; in other words, a $\mathcal{DCEC}^*$ theorem prover is executed, and it attempts to prove or refute ϕ. Naturally, the prover will fail for a lack of starting information, and three things will happen as a result. First, the time is set to t_4 = *"speak1"*. Second, R_3 jumps in the air; this indicates that he has a new message for the human controller. This message is straightforward and honest, and one that can be seen by the human controller after opening the messages window: "I don't know" (Figure 7a). The third thing that happens is that on the AI-side, R_3 is given an additional piece of knowledge:

$$\mathbf{K}(I, t_4, happens(action(I*, \mathbf{S}(I*, t_4, \text{"I don't know"})), t_4)) \quad (1)$$

Formula 1 can be understood as R_3's first-person, or *de se*, knowledge that. at time t_4, he *himself* said "I don't know". The notation used here to capture first-person propositions is drawn from, and the interested reader is pointed to, [9].

A brief clarification re. Formula 1 is necessary here. In order to successfully engineer a solution to the test in the proof described in [29], R_3 must be able to: (1) initiate the action of saying "I don't know" at time t_4; (2) somehow "hear" that it has said "I don't know" at time t_4; and (3) encode the knowledge of what it has heard in such a form that can be reasoned over. Although dialects of $\mathcal{DCEC}^*$ have an operator $\mathbf{P}$ for perception, it is not utlized by the variant of $\mathcal{DCEC}^*$ in the $\mathcal{DCEC}^*$ reasoner used in this paper, which is the same used in [23].

(a) Robot first is ignorant . . . (b) . . . but the robot figures it out.

Fig. 7: R_3 Changing Its Mind

Agent R_3 takes the action of saying "I don't know" at time t_3, and this utterance is simulated by the message displayed as text on the screen (again, pictured in Figure 7a). R_3 then would perceive what he just did, through some combination of auditory sensory input, sensorimotor feedback (e.g. he registers his robotic larynx vibrating as speaking), and other perceptual processes that fuse the relevant sensory input to produce the perception that an utterance has been made. R_3 then concludes[2] that the utterance just perceived was made by either R_3 or some agent that very convincingly sounds like R_3. In short: R_3 perceives that he heard himself say "I don't know" at time t_3. However, in place of formulae containing the perception operator (for reasons just described), we make use of the $\mathbf{S}$ (or "says") operator. The formula thus passed to R_3, meant to simulate the successful completion of this complex perceptual process (the low-level modeling of which is not the focus of this paper), is Formula 1.

The additional knowledge of Formula 1 (= $\mathscr{S}$) is sufficient to allow R_3 to prove ϕ, but it does not by itself trigger the $\mathcal{DCEC}^*$ prover. Thus, very slightly departing from [29], the human controller again enters the same query as before (ϕ followed by a question-mark). Again the $\mathcal{DCEC}^*$ prover is executed, and this time a proof of ϕ is found. R_3 jumps, once again indicating a message, the time is set to t_5 = *"speak2"*, and a message of success is displayed (Figure 7b).

B. Proving Our Solution to the Dumbing Pill Test

The proof of ϕ found by R_3 will now be described. First, the context $\mathbf{\Pi}$, the knowledge which all of the robotic agents start with:

[2]Not in a deliberate inferential sense, but rather in the sense that his perceptual processes through their normal operations met the necessary conditions in order to produce an explicit percept.

$$\forall_{R,t,t_i,t_j \geq t_i,t_k \geq t_i,\psi} \mathbf{C}(\\ t, happens(action(R, ingestDumbPill), t_i) \rightarrow \\ \neg happens(action(R, \mathbf{S}(R, t_j, \psi)))) \quad (2)$$

$$\mathbf{K}(R_3, t_2, ingestDumbPill \oplus ingestPlacebo) \quad (3)$$

$$\forall_t \mathbf{K}(R_3, t, t_1 < t_2, ..., t_4 < t_5) \quad (4)$$

$$\forall_{R,t,p,q} \mathbf{K}(R, t, p \rightarrow q) \wedge \mathbf{K}(R, t, p) \rightarrow \mathbf{K}(R, t, q) \quad (5)$$

$$\forall_{R,t,p,q} \mathbf{K}(R, t, p \rightarrow \neg q) \wedge \mathbf{K}(R, t, q) \rightarrow \mathbf{K}(R, t, \neg p) \quad (6)$$

Formula 2 sets as common knowledge that if a robot ingests a dumbing pill ($ingestDumbPill$), he will not be able to speak after that. Formula 3 simply states that either a dumbing pill or a placebo will be given to robot R_3 at time t_2 (note the symbol $\oplus$ is a shorthand for exclusive-or), while Formula 4 simply relates the discrete moments. Formulae 5 and 6 show that the knowledge of robotic agents are subject to a form of *modus ponens* and *modus tollens*, though note that the form of *modus tollens* chosen for Formula 6 is selected to make inference easier in this particular example. Obviously sophisticated cognitive agents don't carry out proofs from scratch like this, so it would be necessary, longer term, for our ethically correct robots to be in command of *proof methods*: a dedicated class of algorithms pre-enegineered to efficiently generate proofs given minimal input. The "dawn" of the deontic cognitive event calculus, $\mathcal{DCEC}^*$, is the work reported in [30], and the motivated reader can see that even there methods were formalized for the test at hand there (the so-called "false-belief test"), and affirmed as crucial.

Given $\mathbf{\Pi} \cup \{\mathscr{S}\}$ and the $\mathcal{DCEC}^*$ rules of inference, we have sufficient information to prove ϕ, which the reader can verify, and which we have also verified with a RAIR lab-developed $\mathcal{DCEC}^*$prover.

VI. Real-Robot Demonstration

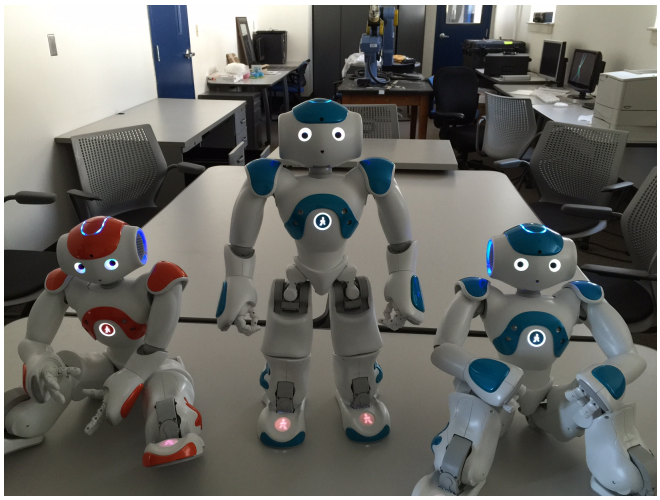


Fig. 8: The Three Aldebaran Naos

The test of robot self-consciousness described above was performed on three Aldebaran Nao humanoid robots, at the RAIR Lab. The simulation transpires as follows:

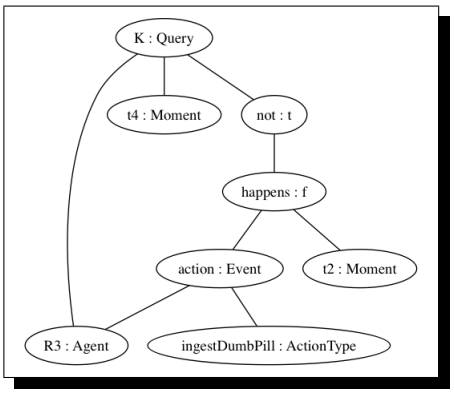


Fig. 9: A $\mathcal{DCEC}^*$ Parse Tree for a Query

1) The robots are programmed to access a $\mathcal{DCEC}^*$ prover, and to interact appropriately with a human tester (corresponding to the aforementioned t_1 = *"apprise"*).
2) In place of physically ingesting pills, the robots are tapped on sensors on their heads (t_2 = *"ingest"*). Unknown to them, two robots have been muted, to simulate being given dumb pills. One robot has not been muted; it was given a placebo.
3) The robots are then asked: "Which pill did you receive?" (t_3 = *"inquire"*), which triggers a query to the $\mathcal{DCEC}^*$ prover. Each robot attempts to prove that it knows, at time t_4, that it did not ingest a dumb pill at time t_2.
4) Each robot fails in this proof attempt, and, accordingly, attempts to report 'I don't know' (t_4 = *"speak1"*). However, two robots, having been muted, are not heard to speak at all. The third robot, however, is able to hear itself speak. It updates its knowledge base to reflect this, and attempts to re-prove the conjecture.
5) This time, it is able to prove the conjecture, and says (t_5 = *"speak2"*) "Sorry, I know now! I was able to prove that I was not given a dumbing pill!"

A video of the demonstration may be found here.[3]

VII. Natural Language Understanding

There is certainly impressive precedent for computational systems able to take English sentences in as input, and yield logicist representations as output. The orthodox form of such representations would be FOL+λ-calculus. We see this approach, accompanied by coverage of the relevant formal terrain, in for example [31] and [32]. Some particularly promising research that follows this line of attack employs logic-based grammar, specifically Combinatory Categorical Grammar (CCG) [33], for parsing natural-language sentences, and then inverse λ-calculus algorithms [34] and other computational-semantics tools. Prominent systems in this vein include: C&Ctools, which uses Curran and Clark's CCG Parser [35]; Bos' computational-semantic Boxer [36];

[3] https://dl.dropboxusercontent.com/u/16443685/NaoBotDemo.mov

and UW SPF [37], that again uses a variant of CCG and various semantic tools.

One specific precedent, relevant to our work, is based on the requirement that the input conform to some controlled subset E' of English, where every processed sentence S is in E'. An example of this precedent is the case of S being in conformity with ACE, and the output being in conformity with **discourse representation theory** (DRT); that is, the output is a **discourse representation structure** (DRS). See, for example, [38], [39], [40]. Another major effort on this front is SemEval-2014 Task [41]: "supervised semantic parsing of Robotic Spatial commands." Here, a specific Robot Commands Treebank was used to train the systems to convert commands in natural language to Robot Control Langauge (RCL).

From the broader perspective of formal, logicist approaches to NLU, we do not subscribe to anything like a Montagovian framework [42], which is model-theoretic in nature. Consistent with what we said earlier in the present paper, our approach to semantics is a proof-theoretic one. Ultimately, therefore, the meaning of natural language is created by the role that formal correlates to sentences in that language play in proofs, or at least in formally specified arguments.

Instead of expressing natural language at the level of FOL, which contra- Blackburn and Bos (2005) we see as severely limiting, we cash out natural language into robust multi-operator quantified intensional logics of unprecedented expressivity.

In this scientific context, the specific NLU technology employed in our demonstration uses a three-step process to convert natural-language questions into $\mathcal{DCEC}^*$ formulae (including formulae serving as queries). The process encompasses syntactic and dependency parsing, semantic analysis, and contextual semantics to generate $\mathcal{DCEC}^*$ formulae. Hence we design a system that skips the use of a dedicated logic-based grammar and instead, directly jump to Wordnet-based [43] semantic analysis. Also, we do not put a constraint on the vocabulary for the query, given the simple nature of the application. However, use of a controlled natural-language subset in congruence with RCL is imminent for more robust and complicated systems.

The query in natural language goes through a number of natural-language pre-processing tools, including POS Tagger [44], Dependency Parser [45], and Word Sense Disambiguation (WSD) [46] tools. Through traversal of the dependency tree generated, we identify the main and auxiliary verbs and their dependencies, and run WSD tools on them. An experimental run through the WSD algorithms implied that Adapted Lesk algorithm [47] is currently the best fit for our present application. We generate a feature vector based on the following features we deemed sufficient for semantic classification of the verbs into the operators of $\mathcal{DCEC}^*$: Perceive, Knowledge, Say, Desire, Intend and Oblige, and action action verbs. This list composes the feature vectors in question:

1) Semantic similarity scores based on WordNet definitions for the senses generated for the verbs to verb senses pertaining to each of the categories mentioned;
2) maximum of the Semantic similarity scores based on WordNet definitions for all the best 3 senses possible for the verb to the verb senses pertaining to categories mentioned. (This is introduced to mask the occasional inaccuracy of the WSD tools.)

A weighted sum of these features is used in construction of an intermediate tree-based logical representation that follows the structure of the dependency tree closely. For further processing, we need inputs from the knowledge-base of the system.

The contextual semantics employed in this NLU system, as mentioned, uses a proof-theoretic approach to generate the final $\mathcal{DCEC}^*$ query. In addition to the knowledge of the robots, the system assumes the following pair of propositions to be true, and uses them to arrive at the $\mathcal{DCEC}^*$ query in the present case:

1) The robot receiving a pill entails ingestion of that pill.
2) The inquirer is looking for the knowledge of the intended respondent at the moment the latter speaks.

Upon receiving the natural language question *Which pill did you receive?*, the NLU system determines that the intended answer will precisely be either the dumb pill or the placebo, and that the listener robot is the agent of Knowledge and Event. In addition, the system uses the knowledge of timestamp of the ingestion of the pill as the moment for the Event and that of the robot speaking as the moment when its knowledge is tested. Hence, using the aforementioned system-wide knowledge, the NLU system generates the following $\mathcal{DCEC}^*$ query, which corresponds to the tree structure shown in Figure 9:

$$\mathbf{K}(R_3, t_4, not(happens(action(R_3, ingestDumbPill), t_2)))$$

VIII. Next Steps

As alert readers have doubtless noted, our robots, whether virtual or physical, are a bit deficient in the NLP direction. Our next step is to introduce the RAIR Lab's semantic NLU system into the equation, so that the pivotal declarative content in $\mathcal{DCEC}^*$ seen above is automatically generated from the English used to dialogue with the robots in question. In addition, the role of self-consciousness, or more precisely the role of *de se* $\mathcal{DCEC}^*$ formulae, within moral reasoning and decision-making, has not yet been fully systematized; this is a second step. There are myriad additional steps that need to ultimately be taken, since of course the goal of engineering morally competent robots is flat-out Brobdingnagian, but one step at a time is the only way forward, and these first two stand immediately before us.

References

[1] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson, "First Steps toward Natural Human-Like HRI," *Autonomous Robots*, vol. 22, no. 4, pp. 411–423, May 2007.

[2] B. F. Malle, S. Guglielmo, and A. Monroe, "Moral, Cognitive, and Social: The Nature of Blame," in *Social Thinking and Interpersonal Behavior*, J. Forgas, K. Fiedler, and C. Sedikides, Eds. Philadelphia, PA: Psychology Press, 2012, pp. 313–331.

[3] P. Bello and M. Guarini, "Introspection and Mindreading as Mental Simulation," in *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, 2010, pp. 2022–2027.

[4] S. Bringsjord, "Psychometric Artificial Intelligence," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 23, no. 3, pp. 271–277, 2011.

[5] S. Bringsjord and B. Schimanski, "What is Artificial Intelligence? Psychometric AI as an Answer," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI–03)*. San Francisco, CA: Morgan Kaufmann, 2003, pp. 887–893. [Online]. Available: http://kryten.mm.rpi.edu/scb.bs.pai.ijcai03.pdf

[6] N. Chapin, B. Szymanski, S. Bringsjord, and B. Schimanski, "A Bottom-Up Complement to the Logic-Based Top-Down Approach to the Story Arrangement Test," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 23, no. 3, pp. 329–341, 2011.

[7] S. Bringsjord and J. Licato, "Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room," in *Foundations of Artificial General Intelligence*, P. Wang and B. Goertzel, Eds. Amsterdam, The Netherlands: Atlantis Press, 2012, pp. 25–47, This url is to a preprint only. [Online]. Available: http://kryten.mm.rpi.edu/Bringsjord_Licato_PAGI_071512.pdf

[8] A. Turing, "Computing Machinery and Intelligence," *Mind*, vol. LIX (59), no. 236, pp. 433–460, 1950.

[9] S. Bringsjord and N. S. Govindarajulu, "Toward a Modern Geography of Minds, Machines, and Math," in *Philosophy and Theory of Artificial Intelligence*, ser. Studies in Applied Philosophy, Epistemology and Rational Ethics, V. C. M¨uller, Ed. New York, NY: Springer, 2013, vol. 5, pp. 151–165. [Online]. Available: http://www.springerlink.com/content/hg712w4l23523xw5

[10] N. S. Govindarajulu, "Towards a Logic-based Analysis and Simulation of the Mirror Test," in *Proceedings of the European Agent Systems Summer School Student Session 2011*, Girona, Spain, 2011. [Online]. Available: http://eia.udg.edu/easss2011/resources/docs/paper5.pdf

[11] L. Floridi, "Consciousness, Agents and the Knowledge Game," *Minds and Machines*, vol. 15, no. 3-4, pp. 415–444, 2005. [Online]. Available: http://www.philosophyofinformation.net/publications/pdf/caatkg.pdf

[12] S. Bringsjord, "Declarative/Logic-Based Cognitive Modeling," in *The Handbook of Computational Psychology*, R. Sun, Ed. Cambridge, UK: Cambridge University Press, 2008, pp. 127–169. [Online]. Available: http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf

[13] ——, "Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness," *Metaphilosophy*, vol. 41, no. 3, pp. 292–312, 2010. [Online]. Available: http://kryten.mm.rpi.edu/sb_on_floridi_offprint.pdf

[14] ——, *What Robots Can and Can't Be*. Dordrecht, The Netherlands: Kluwer, 1992.

[15] ——, "Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline," *Journal of Consciousness Studies*, vol. 14, no. 7, pp. 28–43, 2007. [Online]. Available: http://kryten.mm.rpi.edu/jcsonebillion2.pdf

[16] N. Block, "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences*, vol. 18, pp. 227–247, 1995.

[17] S. Bringsjord, "The Zombie Attack on the Computational Conception of Mind," *Philosophy and Phenomenological Research*, vol. 59, no. 1, pp. 41–69, 1999.

[18] ——, "In Defense of Impenetrable Zombies," *Journal of Consciousness Studies*, vol. 2, no. 4, pp. 348–351, 1995.

[19] S. Bringsjord, R. Noel, and C. Caporale, "Animals, Zombanimals, and the Total Turing Test: The Essence of Artificial Intelligence," *Journal of Logic, Language, and Information*, vol. 9, pp. 397–418, 2000. [Online]. Available: http://kryten.mm.rpi.edu/zombanimals.pdf

[20] S. Bringsjord, N. S. Govindarajulu, S. Ellis, E. McCarty, and J. Licato, "Nuclear Deterrence and the Logic of Deliberative Mindreading," *Cognitive Systems Research*, vol. 28, pp. 20–43, 2014.

[21] S. Bringsjord and N. S. Govindarajulu, "Toward a Modern Geography of Minds, Machines, and Math," *Philosophy and Theory of Artificial Intelligence*, vol. 5, pp. 151–165, 2013.

[22] A. Rao and M. Georgeff, "Modeling Rational Agents within a BDI-Architecture," in *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, 1991, pp. 473–484.

[23] N. Marton, J. Licato, and S. Bringsjord, "Creating and Reasoning Over Scene Descriptions in a Physically Realistic Simulation," in *Proceedings of the 2015 Spring Simulation Multi-Conference*, 2015.

[24] G. Gentzen, "Investigations into Logical Deduction," in *The Collected Papers of Gerhard Gentzen*, M. E. Szabo, Ed. Amsterday, The Netherlands: North-Holland, 1935, pp. 68–131, This is an English version of the well-known 1935 German version.

[25] D. Prawitz, "The Philosophical Position of Proof Theory," in *Contemporary Philosophy in Scandinavia*, R. E. Olson and A. M. Paul, Eds. Baltimore, MD: Johns Hopkins Press, 1972, pp. 123–134.

[26] G. Kreisel, "A Survey of Proof Theory II," in *Proceedings of the Second Scandinavian Logic Symposium*, J. E. Renstad, Ed. Amsterdam, The Netherlands: North-Holland, 1971, pp. 109–170.

[27] N. Francez and R. Dyckhoff, "Proof-theoretic Semantics for a Natural Language Fragment," *Linguistics and Philosophy*, vol. 33, pp. 447–477, 2010.

[28] J. Perry, "The Problem of the Essential Indexical," *Nous*, vol. 13, pp. 3–22, 1979.

[29] S. Bringsjord, "Meeting Floridi's Challenge to Artificial Intelligence from the Knowledge-Game Test for Self-Consciousness," *Metaphilosophy*, vol. 41, no. 3, April 2010.

[30] K. Arkoudas and S. Bringsjord, "Propositional Attitudes and Causation," *International Journal of Software and Informatics*, vol. 3, no. 1, pp. 47–65, 2009. [Online]. Available: http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf

[31] B. Partee, A. Meulen, and R. Wall, *Mathematical Methods in Linguistics*. Dordrecht, The Netherlands: Kluwer, 1990.

[32] P. Blackburn and J. Bos, *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Stanford, CA: CSLI, 2005.

[33] M. Steedman and J. Baldridge, *Combinatory Categorial Grammar*. Wiley-Blackwell, 2005, ch. 5.

[34] A. G. Chitta Baral, Marcos Alvarez Gonzalez, *The Inverse Lambda Calculus Algorithm for Typed First Order Logic Lambda Calculus and Its Application to Translating English to FOL*. Springer, 2012, vol. 7265, ch. 4.

[35] J. R. C. Stephen Clark, "Wide-coverage efficient statistical parsing with ccg and log-linear models," *Computational Linguistics*, vol. 33, no. 4, pp. 493–552, December 2007.

[36] J. B. James R. Curran, Stephen Clark, "Linguistically Motivated Large-Scale NLP with C&C and Boxer," in *Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo)*, 2007, pp. 33–36.

[37] Y. Artzi and L. Zettlemoyer, "Uw spf: The university of washington semantic parsing framework," arXiv:1311.3011, 2013.

[38] N. E. Fuchs, U. Schwertel, and R. Schwitter, "Attempto Controlled English (ACE) Language Manual, Version 3.0," Department of Computer Science, University of Zurich, Zurich, Switzerland, Tech. Rep. 99.03, 1999.

[39] S. Hoefler, "The Syntax of Attempto Controlled English: An Abstract Grammar for ACE 4.0," Department of Informatics, University of Zurich, Zurich, Switzerland, Tech. Rep. ifi-2004.03, 2004.

[40] ——, "The syntax of attempto controlled english: An abstract grammar for ace 4.0," Department of Informatics, University of Zurich, Zurich, Switzerland, Tech. Rep. ifi-2004.03, 2004.

[41] K. Dukes, "Supervised semantic parsing of robotic spatial commands," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 45–53.

[42] R. Montague, "Universal grammar," *Theoria*, vol. 36, no. 3, pp. 373–398, December 1970.

[43] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[44] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research (JMLR)*, 2011.

[45] D. Chen and C. Manning, "A Fast and Accurate Dependency Parser using Neural Networks," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[46] L. Tan, "Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]," https://github.com/alvations/pywsd, 2014.

[47] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2002, vol. 2276, pp. 136–145.