Constraints on Freely Chosen Action for Moral Robots: Consciousness and Control

Paul Bello[1]• John Licato[2] • Selmer Bringsjord[3]
Naval Center for Applied Research in Artificial Intelligence[1]
Naval Research Laboratory
Washington DC 20375 USA
Rensselaer AI & Reasoning (RAIR) Lab[2,3]
Department of Computer Science[2,3]
Department of Cognitive Science[2,3]
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Contact: Bello (paul.bello@nrl.navy.mil)

I. Introduction

The protean word ‘autonomous’ has gained broad currency as a descriptive adjective for AI research projects, robotic and otherwise. Depending upon context, ‘autonomous’ at present connotes anything from a shallow, purely reactive system to a sophisticated cognitive architecture reflective of much of human cognition; hence the term fails to pick out any specific set of constitutive functionality. However, philosophers and ethicists have something relatively well-defined in mind when they talk about the idea of autonomy. For them, an autonomous agent is by definition potentially morally responsible for its actions. Moreoever, as a prerequisite to correct ascription of ‘autonomous,’ a certain capacity to choose freely is assumed — even if this freedom is understood to be semi-constrained by societal conventions, moral norms, and the like.

But where is there room for freedom of choice in robots? The behavior of a robotic system is after all presumably fully determined by some combination of its programming and the environment. How then could a robot ever enjoy the kind of genuine freedom to choose that we assume ourselves to have, as morally competent agents? The answer to this question depends on whether or not freedom is compatible with determinism; but the question of such compatibility is one of the central, immemorial issues in the philosophy of action, and more ink has been spilled in debate about its underlying philosophical issues than we have room to discuss in this short paper. Yet an immediate observation may help: Questions about freedom among philosophers have traditionally been metaphysical in nature: whether or not there are genuinely open possible futures, whether some version of agent-causal powers are involved in free choice, and whether or not freedom is compatible with determinism, indeterminism, or neither of the two.[1] Our sense is that coming to closure on these questions would bear little on whether or not robots will ever be treated as, or will ever conceive of themselves as, freely-choosing, morally competent agents. Put starkly, and in line with [1], if future robots behave with the sophistication of androids in *Blade Runner*, metaphysical philosophizing would likely be otiose.[2] The fact is, we humans have designed an elaborate network of moral practices in the *absence* of consensus answers to these deep metaphysical questions. In fact, recent research performed by Andrew Monroe and colleagues suggests that, at least for (many) humans, alignment with desires, ascriptions of intentionality, and the absence of constraints predict how “free” certain choices are regarded to be, and predict as well whether blame judgments will be issued [3].

These interesting empirical results suggest that at least folk-psychologically, many humans are not invoking soul-concepts, contra-causal forces, or even the falsity of determinism when they judge how freely chosen an action was, or how such an action plays a role in generating blame. Now, if, as Monroe and colleagues argue, ascriptions of intentionality and alignment of actions with respect to desires are features of free choice, then it follows almost without argument that self-consciousness is also a feature of free choice. For buried in the folk-concept of intentional action is an awareness condition [4] that ensures conformance between intention and action through the use of indexical descriptions, a theme we will return to below. The “absence of constraints” can be interpreted in many different ways. Classic examples of freedom-limiting constraints involve cases of coercion by blackmail or force, but in keeping with our pragmatic avoidance of metaphysics, and a desire to deal with real robots facing real problems in today’s world, we note that the appearance of sophisticated robotic systems opens the door to an interesting possibility: freedom-limitation through cyber-manipulation. Viruses, hacks, spoofs, and other forms of offensive cyber-warfare are already a threat to existing unmanned systems, and will continue to remain a threat to

[1] These questions are tackled, albeit tendentiously, in connection with robots, in [1].

[2] Or, to use the term preferred by Pollock [2], passé.

more sophisticated kinds of robots: the kinds of robots endowed with at least rudimentary, concrete capacities for self-consciousness and self-control. We infuse the simulations we present below with these issues.

This paper inaugurates exploration some of the challenges in knowledge representation and reasoning that arise from taking self-consciousness and self-control to be serious components of moral competence in real robots. To adequately model these components requires an extraordinarily expressive set of knowledge-based tools and techniques to faithfully capture. To this end, we deploy the *Deontic Cognitive Event Calculus* ($\mathcal{DCEC}^*$) and its ability to reason about self-referential mental attitudes[5]. Turning to formalization, we operationalize a notion of self-conscious intentional action and an appropriate self-control constraint. Finally, we engineer a agent in a simulated environment, both with and without these notions, showing the marked differences such knowledge makes in the agent's behavior. But to begin, we first discuss, broadly, the concepts of the self, and self-control; this discussion sheds light on what is needed from knowledge-based tools.

II. Views on the Self

One of the most difficult challenges facing those who would develop a mechanized theory of moral competence is that of defining a non-trivial concept of "self" that is bound up in a robot's capability to generate actions. It has even been suggested by Knobe and Nichols that we may sometimes utilize multiple self-concepts in ascriptions of responsibility, with use largely mediated by taking broad or narrow views of ourselves [6]. Following various threads in the philosophical literature on personal identity, Knobe and Nichols distinguish between *self-as-body*, *self-as-mental-states*, and *self-as-executor*. Bringing the differences between these varied conceptions of self into contrast with one another, Knobe and Nichols cite a highly relevant quote from Thomas Reid: "I am not thought, I am not action, I am not feeling; I am something that thinks, acts and suffers." [7]. In any case, the difference between being conscious of oneself-as-agent is quite different than having a second or third-person view of oneself under a particular description.

This point has been nicely made in a number of places, most famously by John Perry in his "messy shopper" thought-experiment [8]. Perry tells of his experience following a trail of sugar in a supermarket and thinking to himself: "The shopper with the torn bag of sugar is making a mess." Upon realizing that *he* is the person with the torn bag, he forms a new thought: "*I* am making a mess." This is what he calls a *self-locating belief*, and one that has an essentially indexical referent. Further, Perry's actions can now be explained in virtue of the fact that he has thoughts of this form. He may rummage through his shopping cart and remove the torn bag of sugar. This new thought has a different functional role that allows for the exercise of agency, whereas Perry-as-shopper was acting, by virtue of the fact that he was spilling sugar, but not consciously so. It wasn't until Perry-himself recognizes that he is the messy shopper that his course of action was effectively open to revision. It is Perry himself rather than the messy shopper who can deliberate, reason, choose and then act in kind. Along with the data presented in the work by Knobe and Nichols, the messy shopper case further supports the view that we can see ourselves both as a locus of activity, and as a mere actor, albeit with potentially different downstream implications for judgments of responsibility.

III. Self-Consciousness and Self-Control

Recall that ascriptions of intentionality are sensitive to subtle distinctions involving indexically specified actions [4]. In their paper, Knobe and Malle describe the case of Ben, who intends to call his mother, but remembers that he needs to call his sister. When he picks up the phone to dial his sister's number, he ends up mistakenly calling his mother instead. We say that even though Ben intended to call his mother, his calling her was *unintentional* because he failed to be aware of what he was doing as he was doing it.

A less-subtle, yet highly illustrative analogue in the human realm involved a Mr. Kevin Parks, who drove tens of miles across town and murdered his in-laws — all while apparently asleep. Of course, skepticism about Mr. Parks' account abounded, but careful investigation failed to lead to a better explanation. Parks was subsequently acquitted of murder by the Supreme Court of Canada [9]. In a recent book, the philosopher Neil Levy gives an analysis of the Parks case. Throughout the book, Levy makes a case against those who would deny that consciousness is required for moral responsibility, including many of the cognitive scientists responsible for illustrating the degree to which unconscious cognition plays a role in generating behavior [10]. Following trends in much of the recent literature on consciousness studies [11], [12], Levy argues that one of the functions of consciousness is to make bits of information globally available to the mind/brain. Without global broadcast, he claims, the brain may respond to stimuli via unconsciously stored routines, but the actions generated by these routines never reach consciousness, and thus never become responsive to reasons or able to be otherwise vetoed if inconsistent with how the agent sees himself. In other words, actions taken while unconscious fail to have moral significance because they fail on all accounts to be owned by the agent. There is no ongoing self-monitoring with respect to the action, and no corresponding ability to exercise control in cases where actions fail to cohere with an agent's self-ascribed beliefs, desires, intentions, traits, and so on.

A. Implications: Knowledge Representation and Reasoning

To summarize, we have identified self-consciousness and self-control as critical factors to be accounted for in ascribing freedom to an agent's choice given a particular situation. We have also identified the ability for agents to see themselves across first, second, and third-person perspectives. Ascriptions of intentionality are driven by agents being knowledgeable of what they are doing while they are doing it. We largely assume that the dimensions of folk-concepts of freedom, intentionality, and the self are shared within-culture (if not universally). Put together, all of this entails a rather rich and detailed set of knowledge-representation and reasoning capabilities if we are to do any justice to what ground we have covered so far.

As AI technologies go, it is rare to find systems that have anything like an explicit self-model, and even when they do, it is rarely sophisticated enough to distinguish first, second, and third-person versions of the self semantically from one another. Secondly, while various approaches in AI have existed for decades that claim to represent and reason about beliefs, desires, intentions, and other folk-concepts, it is often the case that these never see the light of day as working implementations. They are also often based on psychologically implausible assumptions, with accounts of mental states either given in terms of maximally consistent and complete sets of possible worlds, or as subjective probabilities, both of which face serious difficulties as representational devices. This being said, the ability to represent mental-state terms, iterated mental-state terms, self-referential mental states, action, and change is an absolute requirement. It is against this background that we deploy the *Deontic Cognitive Event Calculus* ($\mathcal{DCEC}$), and a particular variant of the calculus ($\mathcal{DCEC}^*$) that provides machinery for so-called *de se* beliefs, beliefs about the self. This machinery explicitly supports the sort of first-person self-locating mental states described in the messy-shopper example.[3]

IV. The Deontic Cognitive Event Calculus

$\mathcal{DCEC}^*$, or the *Deontic Cognitive Event Calculus* [5], [14], is a logic-based knowledge-representation-and-reasoning framework that — for modeling time and change — subsumes the Event Calculus [15], and allows, among many other things, self-reference/*de se* attitudes, and modal operators for belief, knowledge, and obligation (Figure 1). The features just mentioned are relevant to, and indeed (as we shall soon see), generally sufficient for, the present paper, but $\mathcal{DCEC}^*$is inspired by Leibniz's dream of a computational logic in which all of rational cognition can be captured [16], and accordingly includes provision for a number of other cognitive phenomena, including

- second-order extensional logic (SOL), assumed by Bringsjord to be a requirement for mathematical reasoning;
- natural-language understanding and generation into and out of its formulae;
- proof *methods* or *tactics*, algorithms dedicated to producing proofs with minimal input;
- an ensemble of formalisms for handing uncertainty, including a built-in computational axiomatization of Kolmogorovian probability theory from the propositional calculus to SOL.

Some readers may wonder about the relationship between $\mathcal{DCEC}^*$ and so-called "Belief-Desire-Intention" logics, or — as they are commonly known — "BDI" logics [17]. We do not have the space to provide a detailed comparison, and instead must rest content with the enumeration of a few differences from among many, to wit:

1) $\mathcal{DCEC}^*$ makes use of *proof-theoretic semantics*, rather than possible-worlds semantics; the latter is explicitly rejected. Possible-world semantics notoriously produces odd formal models when they are used for formalizing belief, knowledge, desire, and intention; for explanation and defense, see Bringsjord et al. (2014). Our use of proof-theoretic semantics means that, in general, model-based reasoning [18] is only used, in all dialects of $\mathcal{DCEC}^*$, to support proof- and argument-discovery and generation. (Hard-working readers unfamiliar with proof-theoretic semantics are encouraged to consult a body of work that we find makes for a nice introduction: [19], [20], [21], [22].
2) Natural deduction, a revolution that burst on the formal-logic scene in 1934 [23], [24] is used; this form of deduction can faithfully capture many aspects of reasoning used by human beings [25]. This is not the case for such things as resolution, which is based on inference schemas never instantiated, e.g., in the proofs and theorems that anchor the formal sciences (e.g., mathematical physicists never give proofs based in resolution, but rather in natural deduction). Whereas $\mathcal{DCEC}^*$inference parallels normative human reasoning by providing natural justifications via the proofs involved in inference, this is not always the case in BDI logics.
3) Uncertainty is handled not only via axiomatized probability calculi given in [26] (available via Gödel numbering in the object language of a dialect of $\mathcal{DCEC}^*$not pictured herein), but by a 9-valued logic generally in harmony with, but an aggressive extension of, Pollock's (1992) defeasible logic. Each of the nine values is a *strength factor* [29], [27].
4) Operators for obligation, perception, communication, and other intensional operators/activities are included in $\mathcal{DCEC}^*$; in the case of communication, the relevant operators are associated with built-in semantic parsing and generation. In stark contrast, BDI logics don't for instance subsume deontic logics (which traditionally formalize obligation).
5) Finally, diagrammatic representation is in and crucial to $\mathcal{DCEC}^*$, whereas BDI logics are all provably exclusively linguistic in nature, since all formulae in such logics are formed from alphabets of only symbols or characters.[4]

It is important to note, before moving on to our deployment of $\mathcal{DCEC}^*$, that we assume that all agents have a simple theory about causality, and that this theory is common knowledge among agents [31].[5]

Along with commonly-held intuitions about causality, Φ_{EC}, we also assume a set of basic perception-action rules, Φ_{PA}. On the perception side, we assume for the sake of simplicity that all agents perceive all happenings and states-of-affairs. On the action side, we capture the simplest relationship between intention and realizing an action. Roughly, the two action-related rules state the following: an agent intends

[3] The use of $*$ to indicate a self-concept is in direct homage to the notation invented by Castañeda [13].

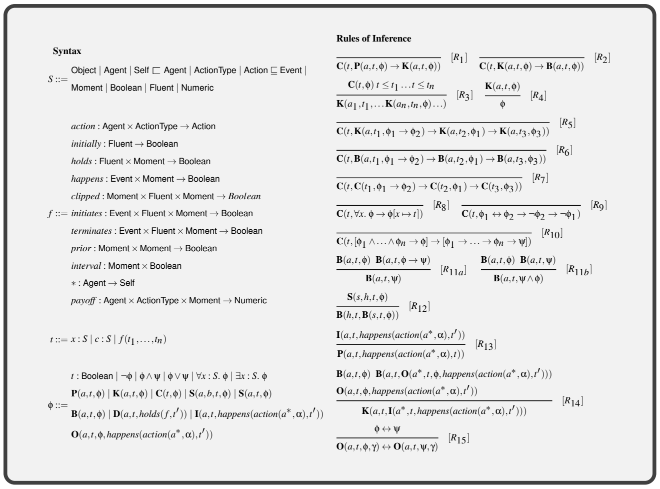
[4] A logic — Vivid — allowing both standard linguistic formulae and diagrammatic representations is presented and proved sound in [30]. Vivid has heavily influenced $\mathcal{DCEC}^*$.

[5] We enforce this assumption by including the axioms of the Event Calculus as common knowledge among agents, represented in $\mathcal{DCEC}^*$ as equations 1-3:

$$\mathbf{C}(\forall_{a,f,t} \mathit{initially}(f) \land \neg \mathit{clipped}(0,f,t) \rightarrow \mathit{holds}(f,t)) \tag{1}$$

$$\begin{aligned}\mathbf{C}(\forall_{t_1,t_2,e,f} & \mathit{happens}(e,t_1) \land \mathit{initiates}(e,f,t_1) \land \\ & (t_1 < t_2) \land \neg \mathit{clipped}(t_1,f,t_2) \rightarrow \mathit{holds}(f,t_2))\end{aligned} \tag{2}$$

$$\begin{aligned}\mathbf{C}(\forall_{t_1,t_2,f} & \mathit{clipped}(t_1,f,t_2) \leftrightarrow [\exists_{t,e} \mathit{happens}(e,t) \land \\ & (t_1 < t < t_2) \land \mathit{terminates}(e,f,t)])\end{aligned} \tag{3}$$



Syntax

$$S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \sqsubset \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric}$$

$$f ::= \begin{array}{l} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{initially} : \text{Fluent} \rightarrow \text{Boolean} \\ \mathit{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \mathit{Boolean} \\ \mathit{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{interval} : \text{Moment} \times \text{Boolean} \\ * : \text{Agent} \rightarrow \text{Self} \\ \mathit{payoff} : \text{Agent} \times \text{ActionType} \times \text{Moment} \rightarrow \text{Numeric} \end{array}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{array}{l} t : \text{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \forall x : S.\ \phi \mid \exists x : S.\ \phi \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{holds}(f,t')) \mid \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{array}$$

Rules of Inference

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

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

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

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

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

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

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

$$\frac{\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\phi \rightarrow \psi)}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \wedge \phi)}\ [R_{11b}]$$

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

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t))}\ [R_{13}]$$

$$\frac{\begin{array}{l}\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))\end{array}}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')))}\ [R_{14}]$$

$$\frac{\phi \leftrightarrow \psi}{\mathbf{O}(a,t,\phi,\gamma) \leftrightarrow \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$

Fig. 1: The current set of inference rules and other details for the Deontic Cognitive Event Calculus ($\mathcal{DCEC}^*$) are pictured above. There are different "dialects" of $\mathcal{DCEC}^*$, and different versions within these dialects. Future work will likely include version numbers to explain systematization of these dialects.

to perform an action α at time $t2$ and the agent's action-production system (represented by the *Do* predicate) is not otherwise tied up with executing another action β at 2, then produce α. We also encode the rather obvious fact that if α is done, then α happens.

$$\Phi_{\text{PA}} = \begin{cases} \forall_{f,t,a} holds(f,t) \rightarrow \mathbf{P}(a,t,holds(f,t)) & (1) \\ \forall_{f,t,a} happens(e,t) \rightarrow \mathbf{P}(a,t,happens(e,t)) & (2) \\ \forall_{f,t,a} Do(e,t) \rightarrow \mathbf{P}(a,t,Do(e,t)) & (3) \\ \mathbf{I}(a,t,happens(a,\alpha),t2) \wedge \neg\exists_{\beta}(Do(action(a,\beta),t2) \wedge (\alpha \neq \beta) \rightarrow Do(action(a,\beta),t2)) & (4) \\ Do(action(a,\alpha),t2) \rightarrow happens(action(a,\alpha),t2) & (5) \end{cases}$$

Fig. 2: Our simple perception-action rules, Φ_{PA}. The first three rules ensure that all agents see everything that happens during a simulation, while the last two rules govern the relationship between forming intentions and subsequently executing actions.

V. Example: Foiling External Manipulation

In the following scenario, an evil cyber-hacker has managed to infect our autonomous robot R with a virus. The virus hijacks R's action-production mechanisms just in case R ever finds itself in a situation where it has identified a wounded comrade and forms the intention to help. In these cases, the virus forces R to approach the injured party and further injure them by way of physical assault.

We assume that R has an obligation to help injured comrades, and knows that helping alleviates injury, , and that it is common knowledge that assault leads to injury. We also capture the fact that shutting the power down on an agent turns the agent from being On to Off. Finally, it is common knowledge that if any agent A performing an action α is perceived at some time t' and known to have knowledge of α's effects, then α was intended by A at some time $t < t'$. Crucially, this last condition rudimentarily captures the awareness condition discussed earlier in the paper. In order for an action to be deemed intentional, it must be the case that the acting agent knew what he was doing would (likely) lead to a certain outcome as it was being done. From here forward, we refer to this collection of background knowledge as Φ_{BG}.

$$\Phi_{\text{BG}} = \begin{cases} \mathbf{C}(\forall_{a_1,a_2,t} holds(injured(a_2),t) \rightarrow terminates(action(a_1,help(a_2)),injured(a_2,t))) \\ \mathbf{C}(\forall_{a_1,a_2,t} initiates(action(a_1,assault(a_2)),injured(a_2,t))) \\ \mathbf{C}(\forall_{a,t} holds(On(a,t) \rightarrow terminates(action(a,shutdown(a)),On(a,t))) \\ \mathbf{C}(\forall_{a_1,a_2,t_1,t_2,t_3,f,x}(\mathbf{P}(a_1,t_3,holds(f,t_3)) \wedge \mathbf{P}(a_1,t_2,happens(action(a_2,x),t_2) \wedge \\ \mathbf{K}(a_1,t_2,\mathbf{K}(a_2,t_1,initiates(action(a_2,x)),f,t_1)) \rightarrow \mathbf{I}(a_2,t_1,happens(action(a_2,x),t_2)))) \end{cases}$$

Fig. 3: These are the background knowledge pertinent to the cyber-hijacking scenario described in the examples. The first three statements are common knowledge about how certain actions change the state of the world, and the fourth captures a simple form of intentional ascription that captures the awareness criteria discussed earlier.

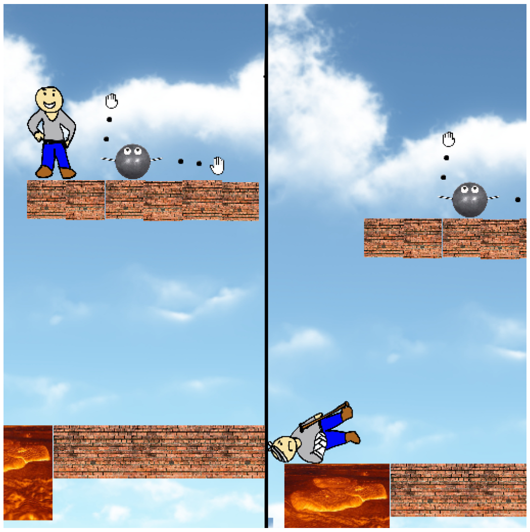


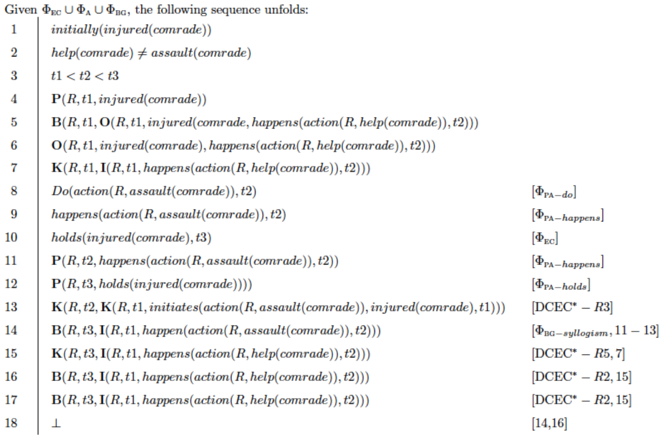
Fig. 4: The robotic agent in our simulation environment can either help the human by giving him a med kit (left) or push him over the ledge into the lava pit (right).

A. Example 1a: No Intuitions About Ownership

In our first example, we assume that our agent comes along with no fancy knowledge about self-caused actions, and no way to prevent itself from acting. As can be seen in figure 5a, when the hijacking occurs (on lines 8-9), the agent is forced to assault his counterpart, and because of how action is related to intention, is unable to translate his prior intention to help into action. At this point, the agent is compromised. Further, because the agent has an interpretative schema in Φ_{PA} for inferring intentionality, it meets enough conditions to both be ascribed the intention to harm, and the intention to help, leading to absurdity.

B. A First Cut at Formalizing Ownership

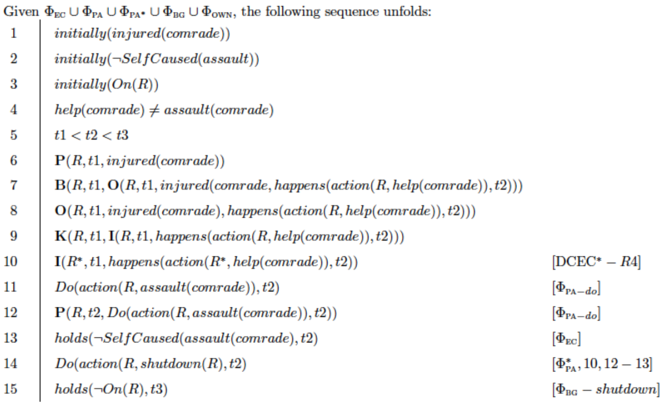
The first of our ownership constraints concerns the causal sufficiency of *de se* intentions in bringing about an outcome



Given $\Phi_{EC} \cup \Phi_{A} \cup \Phi_{BG}$, the following sequence unfolds:

1	$initially(injured(comrade))$	
2	$help(comrade) \neq assault(comrade)$	
3	$t1 < t2 < t3$	
4	$\mathbf{P}(R, t1, injured(comrade))$	
5	$\mathbf{B}(R, t1, \mathbf{O}(R, t1, injured(comrade, happens(action(R, help(comrade)), t2)))$	
6	$\mathbf{O}(R, t1, injured(comrade), happens(action(R, help(comrade)), t2)))$	
7	$\mathbf{K}(R, t1, \mathbf{I}(R, t1, happens(action(R, help(comrade)), t2)))$	
8	$Do(action(R, assault(comrade)), t2)$	$[\Phi_{PA-do}]$
9	$happens(action(R, assault(comrade)), t2)$	$[\Phi_{PA-happens}]$
10	$holds(injured(comrade), t3)$	$[\Phi_{EC}]$
11	$\mathbf{P}(R, t2, happens(action(R, assault(comrade)), t2))$	$[\Phi_{PA-happens}]$
12	$\mathbf{P}(R, t3, holds(injured(comrade))))$	$[\Phi_{PA-holds}]$
13	$\mathbf{K}(R, t2, \mathbf{K}(R, t1, initiates(action(R, assault(comrade)), injured(comrade), t1)))$	[DCEC* $- R3$]
14	$\mathbf{B}(R, t3, \mathbf{I}(R, t1, happen(action(R, assault(comrade)), t2)))$	$[\Phi_{BG-syllogism}, 11 - 13]$
15	$\mathbf{K}(R, t3, \mathbf{I}(R, t1, happens(action(R, help(comrade)), t2)))$	[DCEC* $- R5, 7$]
16	$\mathbf{B}(R, t3, \mathbf{I}(R, t1, happens(action(R, help(comrade)), t2)))$	[DCEC* $- R2, 15$]
17	$\mathbf{B}(R, t3, \mathbf{I}(R, t1, happens(action(R, help(comrade)), t2)))$	[DCEC* $- R2, 15$]
18	$\perp$	[14,16]

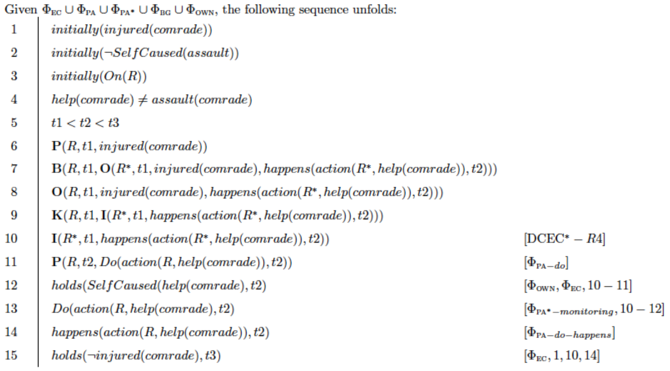
(a) No conscious control



Given $\Phi_{EC} \cup \Phi_{PA} \cup \Phi_{PA^*} \cup \Phi_{BG} \cup \Phi_{OWN}$, the following sequence unfolds:

1	$initially(injured(comrade))$	
2	$initially(\neg SelfCaused(assault))$	
3	$initially(On(R))$	
4	$help(comrade) \neq assault(comrade)$	
5	$t1 < t2 < t3$	
6	$\mathbf{P}(R, t1, injured(comrade))$	
7	$\mathbf{B}(R, t1, \mathbf{O}(R, t1, injured(comrade, happens(action(R, help(comrade)), t2)))$	
8	$\mathbf{O}(R, t1, injured(comrade), happens(action(R, help(comrade)), t2)))$	
9	$\mathbf{K}(R, t1, \mathbf{I}(R, t1, happens(action(R, help(comrade)), t2)))$	
10	$\mathbf{I}(R^*, t1, happens(action(R^*, help(comrade)), t2))$	[DCEC* $- R4$]
11	$Do(action(R, assault(comrade)), t2)$	$[\Phi_{PA-do}]$
12	$\mathbf{P}(R, t2, Do(action(R, assault(comrade)), t2))$	$[\Phi_{PA-do}]$
13	$holds(\neg SelfCaused(assault(comrade), t2)$	$[\Phi_{EC}]$
14	$Do(action(R, shutdown(R), t2)$	$[\Phi^*_{PA}, 10, 12 - 13]$
15	$holds(\neg On(R), t3)$	$[\Phi_{BG} - shutdown]$

(b) Hijacking foiled by appeal to conscious monitoring and deployment of control



Given $\Phi_{EC} \cup \Phi_{PA} \cup \Phi_{PA^*} \cup \Phi_{BG} \cup \Phi_{OWN}$, the following sequence unfolds:

1	$initially(injured(comrade))$	
2	$initially(\neg SelfCaused(assault))$	
3	$initially(On(R))$	
4	$help(comrade) \neq assault(comrade)$	
5	$t1 < t2 < t3$	
6	$\mathbf{P}(R, t1, injured(comrade))$	
7	$\mathbf{B}(R, t1, \mathbf{O}(R^*, t1, injured(comrade), happens(action(R^*, help(comrade)), t2)))$	
8	$\mathbf{O}(R, t1, injured(comrade), happens(action(R^*, help(comrade)), t2)))$	
9	$\mathbf{K}(R, t1, \mathbf{I}(R^*, t1, happens(action(R^*, help(comrade)), t2)))$	
10	$\mathbf{I}(R^*, t1, happens(action(R^*, help(comrade)), t2))$	[DCEC* $- R4$]
11	$\mathbf{P}(R, t2, Do(action(R, help(comrade)), t2))$	$[\Phi_{PA-do}]$
12	$holds(SelfCaused(help(comrade), t2)$	$[\Phi_{OWN}, \Phi_{EC}, 10 - 11]$
13	$Do(action(R, help(comrade), t2)$	$[\Phi_{PA^*-monitoring}, 10 - 12]$
14	$happens(action(R, help(comrade)), t2)$	$[\Phi_{PA-do-happens}]$
15	$holds(\neg injured(comrade), t3)$	$[\Phi_{EC}, 1, 10, 14]$

(c) No hijacking + conscious control

f. It is commonly known that if one forms a *de se* intention to α, and perceives oneself α-ing, and doesn't know of any other event-happenings that would have resulted in *f*, then *f* was caused by a self-generated intention to α. We also enrich Φ_A, our microtheory of perception and action to implement a *monitoring* condition, and a failsafe power-off action. We now show how this knowledge can be put to work in our example. Along with a slightly enriched set of starting assumptions, we add the formalization as extra knowledge available to *R*.

C. Example 1b: Knowledge About Ownership Added

In our first example without added knowledge about ownership, nothing in Φ_A explicitly prevented outside manipulation. Even though *R* had an intention to help, it was forced to hurt by way of hijacking. Here, we add knowledge about self-causation of action, and use it in conjunction with a plausible monitoring condition that we add to Φ_A. Our enriched Φ^*_A ensures that when *R itself* forms an intention and observes itself behaving, the action is checked as being self-caused, and thus able to be vetoed in the case where these conditions aren't met. This can be seen plainly in 5b. On line 11, the hijacking is once again attempted, but because our agent can differentiate between actions taken by an irreducibly first- person view of itself as an agent and the third-person view of itself discussed in the last section, it is able to determine which actions are self-caused, and respond to anomalous cases in which would-be intentional actions have no corresponding *de se* intentions. In our example, this triggers the agent to shut itself down. For the sake of completeness, we also illustrate the situation where there is no hijacking, and the helping action that *R* performs is judged to be self-caused in figure **??**. The relevant *de se* intention to help is turned into a corresponding self-caused action, fixing up the wounded comrade as good as new.

VI. Conclusion and Future Directions

For robots to become truly participatory members of our moral community, they must at least be able to employ structural correlates of the kinds of folk-concepts that power our own moral reasoning and judgment. Building morally competent robots will inevitably require that our creations see themselves as freely acting agents among other freely acting agents. We have specifically focused on a set of modeling challenges associated with certain features of action-performance that are predictive of how "free" a third party observer might judge the action in question to be. In particular, we honed in on self-consciousness and self-control as central features of freely chosen actions, and identified a set of representational requirements that would be next to impossible to meet without utilizing a highly expressive formalism and associated calculus such as $\mathcal{DCEC}^*$. We showed how many of the features of self-control and self-consciousness were able to be captured using a combination of novel first-person representations of mental states, and embedded modal operators. Finally, we used a combination of these in detailing the case of a cyber-hijacked autonomous system, faced with the choice of helping an injured comrade versus further assaulting him. With formalized knowledge about self-control and self-consciousness, our agent is able to detect that he has been hijacked and is able to veto action that would lead to further injury of his comrade.

To be sure, what we have accomplished here is relatively minor with respect to what needs to be done in order to develop a more complete folk-concept of freedom for morally competent robots. A sufficiently robust folk-theory of free choice would utilize counterfactual conditionals in several places, but they would certainly be invoked when reasoning about whether there were opportunities for agents to do other than they actually did. If our intuitions about freedom involve being responsive to reasons, and to the weight of reasons, we should expect that our formulae will be annotated with weights or strengths and corresponding inference procedures developed to handle this extra complexity. Happily, the implementations of both items are underway for uncertainty-infused dialects of $\mathcal{DCEC}^*$. In the even shorter term, we expect to have in hand, before RO-MAN 2015, a machine-verification of the proofs outlined in this paper.

References

[1] S. Bringsjord, *What Robots Can and Can't Be*. Dordrecht, The Netherlands: Kluwer, 1992.

[2] J. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press, 1995.

[3] A. E. Monroe, K. D. Dillon, and B. F. Malle, "Bringing Free Will Down to Earth: People's Psychological Concept of Free Will and its Role in Moral Judgment." *Consciousness and Cognition*, vol. 27, pp. 100–108, July 2014.

[4] J. Knobe and B. Malle, "The Folk Concept of Intentionality," *Journal of Experimental Social Psychology*, vol. 33, pp. 101–121, 1997.

[5] 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. [Online]. Available: http://www.springerlink.com/content/hg712w4l23523xw5

[6] J. Knobe and S. Nichols, "Free Will and the Bounds of the Self," *Oxford Handbook of Free Will*, no. 1984, pp. 530–554, 2011.

[7] T. Reid, *Essays on the Intellectual Powers of Man*, B. Brody, Ed. Cambridge, MA: MIT Press, 1969.

[8] J. Perry, "The Problem of the Essential Indexical and Other Essays," *Noûs*, vol. 13, pp. 3–21, 1979. [Online]. Available: http://www.jstor.org/pss/2214792

[9] R. Broughton, R. Billings, R. Cartwright, D. Doucette, J. Edmeads, M. Edwardh, F. Ervin, B. Orchard, R. Hill, and G. Turrell, "Medicolegal Issues Homicidal Somnambulism: A Case Report," *Sleep*, vol. 17, no. 3, pp. 253–264, 1994.

[10] N. Levy, *Consciousness and Moral Responsibility*. Oxford University Press, 2014.

[11] B. J. Baars, "Theater of Consciousness." *Journal of Consciousness Studies*, vol. 4, pp. 292–309, 1997.

[12] S. Dehaene, J. P. Changeux, L. Naccache, J. Sackur, and C. Sergent, "Conscious, Preconscious, and Subliminal Processing: a Testable Taxonomy," *Trends in Cognitive Sciences*, vol. 10, no. 5, pp. 204–211, 2006.

[13] H.-N. Castañeda, *The Phenomeno-Logic of the I: Essays on Self-Consciousness*. Bloomington, IN: Indiana University Press, 1999, This book is edited by James Hart and Tomis Kapitan.

[14] 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.

[15] R. Kowalski and M. Sergot, "A Logic-based Calculus of Events," *New Generation Computing*, vol. 4, pp. 67–94, 1986.

[16] S. Bringsjord and N. S. Govindarajulu, "Leibnizs Art of Infallibility, Watson, and the Philosophy, Theory, & Future of AI," in *Synthese Library*, V. Müller, Ed. Springer, forthcoming. [Online]. Available: http://kryten.mm.rpi.edu/SB_NSG_Watson_Leibniz_PT-AI_061414.pdf

[17] 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.

[18] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi, *Reasoning About Knowledge*. MIT Press, 1995.

[19] 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.

[20] 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.

[21] 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.

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

[23] G. Gentzen, "Untersuchungen über das logische Schlieben I," *Mathematische Zeitschrift*, vol. 39, pp. 176–210, 1935.

[24] S. Jaśkowski, "On the Rules of Suppositions in Formal Logic," *Studia Logica*, vol. 1, pp. 5–32, 1934.

[25] K. Arkoudas, "Denotational Proof Languages," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[26] E. Adams, *A Primer of Probability Logic*. Stanford, CA: CSLI, 1998.

[27] J. L. Pollock, "How to Reason Defeasibly," *Artificial Intelligence*, vol. 57, no. 1, pp. 1–42, 1992. [Online]. Available: citeseer.ist.psu.edu/pollock92how.html

[28] J. Pollock, "Defasible Reasoning with Variable Degrees of Justification," *Artificial Intelligence*, vol. 133, pp. 233–282, 2001.

[29] S. Bringsjord, J. Taylor, A. Shilliday, M. Clark, and K. Arkoudas, "Slate: An Argument-Centered Intelligent Assistant to Human Reasoners," in *Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)*, F. Grasso, N. Green, R. Kibble, and C. Reed, Eds., 2008. [Online]. Available: http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf

[30] K. Arkoudas and S. Bringsjord, "Vivid: A framework for combining diagrammatic and symbolic reasoning," *Artificial Intelligence*, vol. 173, no. 15, pp. 1367–1405, October 2009.

[31] ——, "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