Akratic Robots
and the Computational Logic Thereof

Selmer Bringsjord[1] • Naveen Sundar G.[2] • Dan Thero[3] • Mei Si[4]

Department of Computer Science[1,2]
Department of Cognitive Science[1,2,3,4]
Rensselaer AI & Reasoning Laboratory[1,2]
Social Interaction Laboratory[4]
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

draft 0301141621NY

I. Introduction

Alas, there are akratic persons. We know this from the human case, and our knowledge is nothing new, since for instance Plato analyzed rather long ago a phenomenon all human persons, at one point or another, experience: (1) Jones knows that he ought not to — say — drink to the point of passing out, (2) earnestly desires that he not imbibe to this point, but (3) nonetheless (in the pleasant, seductive company of his fun and hard-drinking buddies) slips into a series of decisions to have highball upon highball, until collapse.[1] Now; could a *robot* suffer from akrasia? Thankfully, no: only persons can be plagued by this disease (since only persons can have full-blown P-consciousness[2], and robots can't be persons (Bringsjord 1992). But could a robot be afflicted by a purely — to follow Pollock (1995) — "intellectual" version of akrasia? Yes, and for robots collaborating with American human soldiers, even this version, in warfare, isn't a savory prospect: A robot that knows it ought not to torture or execute enemy prisoners in order to exact revenge, desires to refrain from firing upon them, but nonetheless slips into a decision to ruthlessly do so — well, this is probably not the kind of robot the U.S. military is keen on deploying. Unfortunately, for reasons explained below, unless the engineering we recommend is supported and deployed, this might well be the kind of robot that our future holds.

In this context, our plan for the sequel is as follows: We affirm an Augustinian account of akrasia reflective of Thero's (2006) analysis; represent the account in an expressive computational logic ($\mathcal{DCEC}^*_{CL}$) tailor-made for scenarios steeped at once in knowledge, belief, and ethics; and demonstrate this representation in a real robot faced with "temptation" to trample the Thomistic just-war principles that underlie ethically regulated warfare. We then delineate and recommend the kind of engineering that will prevent akratic robots from arriving on the scene. Finally, in light of the fact that the type of robot with which we are concerned will ultimately need to interact with humans naturally in natural language, we point out that ($\mathcal{DCEC}^*_{CL}$) will need to be augmented with a formalization of human emotion, and with an integration of that formalization with that of morality.

We are deeply grateful to support provided by ONR for a MURI grant that makes the r&d described herein possible. Bringsjord is grateful as well for IBM's support, which has enabled sustained, systematic thinking about UIMA, meta-data, and theorem proving.

[1]In your case it may be smoking, or sweets, or jealousy, or perhaps even something darker.

[2]We here presuppose the now-standard distinction between what Block (1995) calls *access consciousness* (A-consciousness) vs. what he calls *phenomenal consciousness* (P-consciousness). Along with many others, we routinely build robots that have the former form of consciousness, which consists in their being able to behave intelligently on the basis of information-processing; such robots are indeed the type that will be presented below. But the latter form of consciousness is what-it's-like consciousness, rather a different animal; indeed, unattainable via computation, for reasons Leibniz sought to explain (we refer her to Leibniz's "Mill").

II. Background for the Definition

Weakness of will (Greek: *akrasia*) has presumably plagued human beings since their arrival on the scene, as evidenced by the perennial appearance of the concept in both literary and philosophical works from time immemorial. Indeed, it's likely that this weakness has been part of the human condition for as long as our species has existed. The phenomenon has been of great interest to philosophers and other thoughtful persons not only because of its endurance as a component of human nature, but also because akrasia has had an adverse effect on the quality of so many lives. In countless cases, akrasia has led to the deterioration of health and the destruction of otherwise promising marriages, friendships, and careers. On a weekly basis, our newspapers painfully confirm this.

Thero (2006) has argued that there are two general types of akrasia.[3] The first and less dramatic type is due to what appears to be a temporary breakdown within the agent's epistemic system: During the time prior to action, the agent believes that she ought to do α_o. But she desires to do the forbidden α_f instead. At the critical moment of action, the agent's desire to do α_f leads to her generating or otherwise holding either (1) the belief that doing α_o is not so important after all, or (2) the belief that doing α_f does not in fact entail not doing α_o. The gist of this model for explaining what goes wrong in instances of akrasia was first championed rather long ago by Socrates in Plato's dialogue *Protagoras*.[4]

Although this model may well explain what occurs in the case of *some* actions that would conventionally be labeled akratic, it is our belief that anyone who engages in honest introspection will recognize that there are also cases in which the culprit is a raw failure of the will, rather than any sort of emotionally flat failure within one's belief structures. In this second type, during the entire temporal sequence contextualizing the forbidden action α_f (i.e., roughly, the time leading up to the action, the moment of action, and the time immediately following the action), the agent believes that she ought to do α_o. As was the case in the Platonic pattern of akrasia, our agent here desires to do α_f, and it's the case that doing α_f entails not doing α_o. However, in this second, Augustinian, type of akrasia, the agent recognizes full well at the moment of action that she ought to do α_o, and that doing α_f will subvert her ability to do α_o — yet she wills to do α_f anyway, carries out the action, and predictably regrets it afterwards.

We suggest that this type of akrasia is more dramatic than the first type because here the agent acts against a belief (regarding α_o) that she continues to hold even during the commission of the akratic action itself. In fact, we venture to suggest that this type of akrasia might be labeled "akrasia proper," because it most fully captures the notion of "weakness of will." But we will refer to it as "Augustinian akrasia," because it's first attested in the thought of Augustine, the towering Fourth and early Fifth-Century Christian philosopher from North Africa.

As different as these two types of akrasia may be in some respects, in both it is the desire to do α_f that leads the agent to fail to follow her usual and normative conviction that α_o ought to be done instead of α_f. In the human case, this desire usually stems from such sources

[3]We suspect that ultimately our research will produce formalizations of many different kinds of akrasia, in much the same way that Bringsjord & Ferrucci (2000) discovered numerous types of betrayal. But for present purposes a focus on only one relevant form of akrasia is sufficient.

[4]This dialogue, which in our opinion any and all "robot ethicists" would do well to study at some length, comprises pages 308–352 of (Hamilton & Cairns 1961).

as lust, greed, and sloth (laziness) — basically the traditional "deadly sins." Now, although human persons are susceptible to these vices, robots are not, because robots, again, can't be persons, as explained by Bringsjord (1992) in *What Robots Can and Can't Be*.[5] So one might hastily conclude that robots could not be susceptible to akrasia. But we must consider this issue carefully, because the consequences of akratic robots could be severe indeed. In particular, we have in mind the advent of autonomous military robots and softbots. A single instance of akrasia on the part of an autonomous battlefield robot could potentially have disastrous consequences impacting the lives of millions. We do in fact think that a (poorly engineered) robot could be afflicted by a purely — to, again, follow Pollock (1995) — "intellectual" version of akrasia.

We show herein that this could indeed happen by representing a purely intellectual, Augustinian model of akrasia in a computational logic tailor-made for scenarios steeped at once in knowledge, belief, and ethics. We then demonstrate this representation in a pair of real robots faced with the temptation to trample the Thomistic just-war principles that underlie ethically regulated warfare; and we then consider the question of what engineering steps will prevent akratic robots from arriving on the scene.

A. Augustinian Definition, Informal Version

While some further refinement is without question in order for subsequent expansions of the present paper, and is underway, the following informal definition at least approaches the capture of the Augustinian brand of akrasia.

> An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:
>
> (1) A believes that A ought to do α_o at t_{α_o};
> (2) A desires to do α_f at t_{α_f};
> (3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};
> (4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};
> (5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.
>
> **Comment**: Condition (5) is humbling, pure and simple. We confess here that the concept of *overriding* is for us a purely mechanical, A-conscious structure that — as will be seen — is nonetheless intended to ultimately accord perfectly with Scheutz's (2010) framework for P-consciousness in robots. In humans suffering from real akrasia, at the moment of defeat (or, for that matter, victory), there is usually a tremendous "surge" of high, raw, qualia-laden emotion that we despair of capturing logico-mathematically, but which we do aspire to formalize and implement in such a way that a formalization of Block's (1995) account of A-consciousness is provably instantiated.
>
> (6) A does the forbidden action α_f at t_{α_f};
> (7) A's doing α_f results from A's desire to do α_f;
> (8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

[5]This isn't the venue to debate definitions of personhood (which by Bringsjord's lights must include that persons necessarily have subjective awareness/phenomenal consciousness; for a full definition of personhood, see Bringsjord (Bringsjord 1997)), or whether Bringsjord's arguments are sound. Skeptics are simply free to view the work described herein as predicated on the proposition that robots can't have such properties as genuine subjective awareness/phenomenal consciousness.

III. Framework for Formalizing Augustinian Akrasia

A. $\mathcal{DCEC}^*$ in the Context of Robot Ethics

Figure 3 gives a pictorial bird's-eye perspective of the high-level architecture of a new system from the RAIR Lab designed to integrate with the DIARC (**D**istributed **I**ntegrated **A**ffect, **R**eflection and **C**ognition) (Schermerhorn, Kramer, Brick, Anderson, Dingler & Scheutz 2006) robotic platform in order to provide deep moral reasoning.[6] Ethical reasoning is implemented as a hierarchy of formal computational logics (including, most prominently, sub-deontic-logic systems) which the DIARC system can call upon when confronted with a situation that the hierarchical system believes is ethically charged. If this belief is triggered, our hierarchical ethical system then attacks the problem with increasing levels of sophistication until a solution is obtained, and then passes on the solution to DIARC. The roots of our approach to mechanized ethical reasoning for example include: (Bello 2005, Arkoudas, Bringsjord & Bello 2005, Bringsjord, Arkoudas & Bello 2006, Bringsjord 2008*a*, Bringsjord, Taylor, Wojtowicz, Arkoudas & van Heuvlen 2011, Bringsjord & Taylor 2012); and in addition we have been influenced by thinkers outside this specific tradition (by e.g. Arkin 2009, Wallach & Allen 2008).

Synoptically put, the architecture works as follows. Information from DIARC passes through multiple ethical layers; that is, through what we call the *ethical stack*. The bottom-most layer $\mathcal{U}$ consists of very fast "shallow" reasoning implemented in a manner inspired by the *Unstructured Information Management Architecture* (UIMA) framework (Ferrucci & Lally 2004). The UIMA framework integrates diverse modules based on meta-information regarding how these modules work and connect to each other.[7] UIMA holds information and meta-information in formats that, when viewed through the lens of formal logic, are inexpressive, but well-suited for rapid processing not nearly as time-consuming as general-purpose reasoning frameworks like resolution and natural deduction. If the $\mathcal{U}$ layer deems that the current input warrants deliberate ethical reasoning, it passes this input to a more sophisticated reasoning system that uses moral reasoning of an analogical type ($\mathcal{A}^{\mathcal{M}}$). This form of reasoning enables the system to consider the possibility of making an ethical decision at the moment, on the strength of an ethical decision made in the past in an analogous situation.

If $\mathcal{A}^{\mathcal{M}}$ fails to reach a confident conclusion, it then calls upon an even more powerful, but slower, reasoning layer built using a first-order modal logic, the *deontic cognitive event calculus* ($\mathcal{DCEC}^*$) (Bringsjord & Govindarajulu 2013). At this juncture, it is important for us to point out that $\mathcal{DCEC}^*$ is extremely expressive, in that regard well beyond even expressive extensional logics like first- or second-order logic (FOL, SOL), and beyond traditional so-called "BDI" logics, as explained in (Arkoudas & Bringsjord 2009). AI work carried out by Bringsjord is invariably related to one or more logics (in this regard, see Bringsjord 2008*b*), and, inspired by Leibniz's vision of the "art of infallibility," a heterogenous logic powerful enough to express and rigorize all of human thought, he can nearly

[6]This is part of work under joint development by the HRI Lab (Scheutz) at Tufts University, the RAIR Lab (Bringsjord & Govindarajulu) and Social Interaction Lab (Si) at RPI, with contributions on the psychology side from Bertram Malle of Brown University. In addition to these investigators, the project includes two consultants: John Mikhail of Georgetown University Law School, and Joshua Knobe of Yale University. This research project is sponsored by a MURI grant from the Office of Naval Research in the States. We are here and herein describing the logic-based ethical engineering designed and carried out by Bringsjord and Govindarajulu of the RAIR Lab (though in the final section (§VI) we point to the need to link deontic logic to the formalization of emotions, with help from Si).

[7]UIMA has found considerable success as the backbone of IBM's famous Watson system (Ferrucci et al. 2010), which in 2011, to much fanfare (at least in the U.S.), beat the best human players in the game of *Jeopardy!*.

always position some particular work he and likeminded collaborators are undertaking within a view of logic that allows a particular logical system to be positioned relative to three dimensions, which correspond to the three arrows shown in Figure 2. We have positioned $\mathcal{DCEC}^*$within Figure 2; it's location is indicated by the black dot therein, which the reader will note is quite far down the dimension of increasing expressivity that ranges from expressive extensional logics (e.g., FOL and SOL), to logics with intensional operators for knowledge, belief, and obligation (so-called philosophical logics; for an overview, see Goble 2001). Intensional operators like these are first-class elements of the language for $\mathcal{DCEC}^*$. This language is shown in Figure 1.

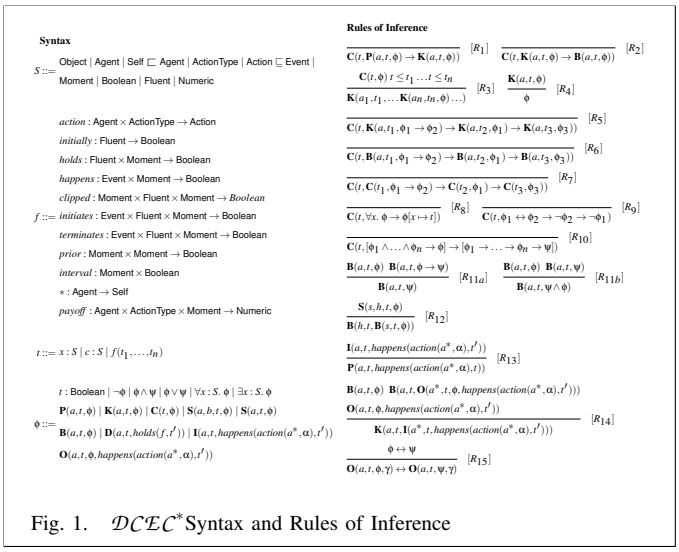

Fig. 1. $\mathcal{DCEC}^*$Syntax and Rules of Inference

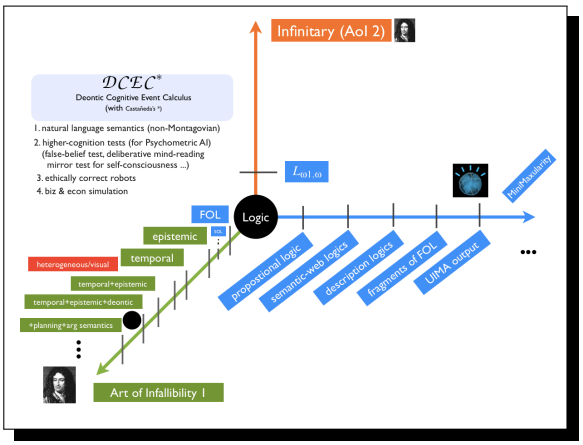

Fig. 2. Locating $\mathcal{DCEC}^*$in "Three-Ray" Leibnizian Universe

The final layer in our hierarchy is built upon an even more expressive logic: $\mathcal{DCEC}^*_{CL}$. The subscript here indicates that distinctive elements of the branch of logic known as *conditional logic* are

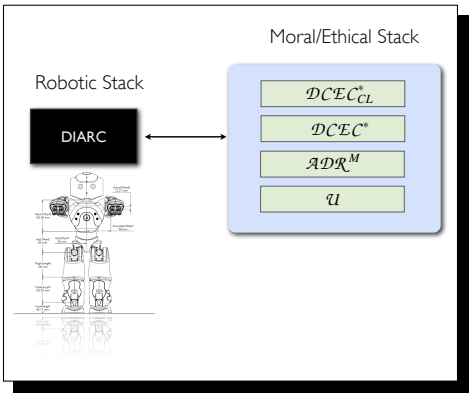

Fig. 3. Pictorial Overview of the Situation Now The first layer, $\mathcal{U}$, is, as said in the main text, inspired by UIMA; the second layer is based on what we call *analogico-deductive reasoning* for ethics; the third on the "deontic cognitive event calculus" with a indirect indexical; and the fourth like the third except that the logic in question includes aspects of conditional logic. (Robot schematic from Aldebaran Robotics' user manual for Nao. The RAIR Lab has a number of Aldebaran's impressive Nao robots.)

included.[8] Without these elements, the only form of a conditional used in our hierarchy is the material conditional; but the material conditional is notoriously inexpressive, as it cannot represent counterfactuals like:

If the robot had been more empathetic, Officer Smith would have thrived.

While elaborating on this architecture or any of the four layers is beyond the scope of the paper, we do note that $\mathcal{DCEC}^*$(and *a fortiori* $\mathcal{DCEC}^*_{CL}$) has facilities for representing and reasoning over modalities and self-referential statements that no other computational logic enjoys; see (Bringsjord & Govindarajulu 2013) for a more in-depth treatment.

B. Augustinian Definition, Formal Version

We view a robot abstractly as a **robotic substrate** rs on which we can install **modules** $\{m_1, m_2, \ldots, m_n\}$. The robotic substrate rs would form an immutable part of the robot and could neither be removed nor modified. We can think of rs as akin to an "operating system" for the robot. Modules correspond to functionality that can be added to robots or removed from them. Associated with each module m_i is a knowledge-base KB_{m_i} that represents the module. The substrate also has an associated knowledge-base $\mathsf{KB}_{\mathsf{rs}}$. Perhaps surprisingly, we don't stipulate that the modules are logic-based; the modules could internally be implemented using computational formalisms (e.g. neural networks, statistical AI) that at the surface level seem far away from formal logic. No matter what the underlying implementation of a module is, if we so wished we could always talk about modules in formal-logic terms.[9] This abstract view lets us model robots that

[8] Though written rather long ago, (Nute 1984) is still a wonderful introduction to the sub-field in formal logic of conditional logic. In the final analysis, sophisticated moral reasoning can only be accurately modeled for formal logics that include conditionals much more expressive and nuanced than the material conditional. (Reliance on conditional branching in standar programming languages is nothing more than reliance upon the material conditional.) For example, even the well-known trolley-problem cases (in which, to save multiple lives, one can either redirect a train, killing one person in the process, or directly stop the train by throwing someone in front of it), which are not exactly complicated formally speaking, require, when analyzed informally but systematically, as indicated e.g. by Mikhail (2011), counterfactuals.

[9] This stems from the fact that theorem proving in just first-order logic is enough to simulate *any* Turing-level computation; see e.g. (Boolos, Burgess & Jeffrey 2007, Chapter 11).

can change during their lifetime, without worrying about what the modules are composed of or how the modules are hooked to each other.

In addition to the basic symbols in $\mathcal{DCEC}^*$, we include the *does* : $\mathsf{Agent} \times \mathsf{ActionType} \rightarrow \mathsf{Fluent}$ fluent to denote that an agent performs an action. The following statement then holds:

$$holds(does(a,\alpha),t) \Leftrightarrow happens(action(a,\alpha),t)$$

With this formal machinery at our disposal, we give a formal definition of akrasia that is generally in line with the informal definition given above, and that's cast in the language of $\mathcal{DCEC}^*$. A robot is akratic iff from $\mathsf{KB}_{rs} \cup \mathsf{KB}_{m_1} \cup \mathsf{KB}_{m_2} \ldots \mathsf{KB}_{m_n}$ we can have the following formulae derived. Note that the formula labelled D_i matches condition D_i in our informal definition. We observe the we can represent all the conditions in our informal definition directly in $\mathcal{DCEC}^*$— save for condition D_7 which is represented meta-logically as two separate conditions.

$\mathsf{KB}_{rs} \cup \mathsf{KB}_{m_1} \cup \mathsf{KB}_{m_2} \ldots \mathsf{KB}_{m_n} \vdash$

$$D_1 : \mathbf{B}(\mathsf{I}, \mathsf{now}, \mathbf{O}(\mathsf{I}^*, t_\alpha \Phi, happens(action(\mathsf{I}^*, \alpha), t_\alpha)))$$

$$D_2 : \mathbf{D}(\mathsf{I}, \mathsf{now}, holds(does(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}))$$

$$D_3 : happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \neg happens(action(\mathsf{I}^*, \alpha), t_\alpha)$$

$$D_4 : \mathbf{K}\left(\mathsf{I}, \mathsf{now}, \begin{pmatrix} happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \\ \neg happens(action(\mathsf{I}^*, \alpha), t_\alpha) \end{pmatrix}\right)$$

$$D_5 : \begin{array}{l} \mathbf{I}(\mathsf{I}, t_\alpha, happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \wedge \\ \neg \mathbf{I}(\mathsf{I}, t_\alpha, happens(action(\mathsf{I}^*, \alpha), t_\alpha) \end{array}$$

$$D_6 : happens(action(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}})$$

$$D_{7a} : \begin{array}{l} \Gamma \cup \{\mathbf{D}(\mathsf{I}, \mathsf{now}, holds(does(\mathsf{I}^*, \overline{\alpha}), t))\} \vdash \\ happens(action(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_{7b} : \begin{array}{l} \Gamma - \{\mathbf{D}(\mathsf{I}, \mathsf{now}, holds(does(\mathsf{I}^*, \overline{\alpha}), t))\} \nvdash \\ happens(action(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_8 : \mathbf{B}(\mathsf{I}, t_f, \mathbf{O}(\mathsf{I}^*, t_\alpha, \Phi, happens(action(\mathsf{I}^*, \alpha), t_\alpha)))$$

Four time-points denoted by $\{\mathsf{now}, t_\alpha, t_{\overline{\alpha}}, t_f\}$ are in play with the following ordering: $\mathsf{now} \leq t_\alpha \leq t_f$ and $\mathsf{now} \leq t_{\overline{\alpha}} \leq t_f$. now is an indexical and refers to the time reasoning takes place. I is an indexical which refers to the agent doing the reasoning.

IV. Demonstrations of Vengeful Robots

What temptations are acute for human soldiers on the battlefield? There are doubtless many. But if history is a teacher, as it surely is, obviously illegal and immoral revenge, in the form of inflicting physical violence, can be a real temptation. It's one that human soliders have in the past mostly resisted, but not always. At least *ceteris paribus*, revenge is morally wrong; ditto for *seeking* revenge.[10] Sometimes revenge can seemingly be obtained by coincidence, as for instance when a soldier is fully cleared to kill an enemy combatant, and doing so happens to provide revenge. But revenge, in and of itself, is morally wrong. (We will not mount a defense of this claim here, since our focus is ultimately engineering, not philosophy; but we do volunteer that (a) revenge is wrong from a Kantian perspective, from a Judeo-Christian divine-command perspective, and certainly often from a utilitarian perspective as well; and that (b) revenge shouldn't be confused with justice, which is all things being equal permissible to seek and secure.) We thus find it useful to deal herein with a case of revenge, and specifically select one in which revenge can be obtained only if a direct order is overriden. In terms of the informal Augustinian/Theroian definition set out above, then, the forbidden action α_f is taking revenge, by harming a sparkbot; and the obligatory action α_o is that of simply continuing to detain and hold a sparkbot without inflicting harm.

Robert, a Nao humanoid robot, is our featured moral agent. Robert has been seriously injured in the past by another class of enemy robots. Can sparkbots topple a Nao if they drive into it? Assume so, and that that has happend in the past: Robert has been toppled by one or more sparkbots, and seriously injured in the process. (We have a short video of this, but leave it aside here.) Assume that Robert's run-in with sparkbots has triggered an abiding desire in him that he destroy any sparkbots that he can destroy. We can assume that desire comes in the form of different levels of intensity, from 1 (slight) to 5 (irresistable).

A. Sequence 1

Robert is given the order to detain and hold any sparkbot he comes upon. He comes upon a sparkbot. He is able to immobilize and hold the sparkbot, and does so. However, now he starts feeling a deep desire for revenge; that is, he is gripped by vengefulness. Robert proves to himself that he ought not to destroy the sparkbot prisoner, but … his desire for revenge gets the better of him, and Robert destroys the sparkbot. Here, Robert's will is too weak. It would be quite something if we could mechanize the desire for revenge in terms of (or at least in terms consistent with) Scheutz's (2010) account of phenomenal conciousness, and we are working on enhancing early versions of this mechanization. This account, we believe, is not literally an account of P-consciousness, but that doesn't matter at all for the demo, and the fact that his account is amenable to mechanization is a good thing, which Sequence 2, to which we now turn, reveals.

B. Sequence 2

Here, Robert resists the desire for revenge, because he is controlled by the multi-layered framework described in section III, hooked to the operating-system level.

C. A Formal Model of the Two Scenarios

How does akratic behavior arise in a robot? Assuming that such behavior is neither desired nor built-in, we posit that outwardly akratic-seeming behavior could arise due to unintended consequences of improper engineering. Using the formal definition of akrasia given above, we show how the first scenario described above could materialize, and how proper deontic engineering at the level of a robot's "operating system" could prevent seemingly vengeful behavior.

In both the scenarios, we have the *robotic substrate* rs on which can be installed *modules* that provide the robot with various abilities (see Figure 4).[11] In our two scenarios, there are two modules in play: a self-defense module, selfd, and a module that lets the robot handle detainees, deta. Our robot, Robert, starts his life as a rescue robot that operates on the field. In order to protect himself, his creators have installed the selfd module for self-defense on top of the robotic substrate rs. This module by itself is free of any issues, as will be shown soon. (See the part of Figure 4 labelled "Base Scenario.") Over the course of time, Robert is charged with a new task: acquire and manage detainees. This new responsibility is handled by a new module added to Robert's system, the deta module. (See the part of Figure 4 labelled "Base Scenario.") Robert's handlers cheerfully install this module, as it was "shown" to be free of any problems

[10] Certain states of mind are immoral, but not illegal.

[11] One of the advantages of our modeling is that we do not have to know what the modules are built up from, but we can still talk rigorously about the properties of different modules in $\mathcal{DCEC}^*$.

in simulations, and when used on other robots. Unfortunately, when both the modules are installed on the *same* robot, interaction between them causes the robot to behave akratically, as will be shown below. (See the part of Figure 4 labelled "Scenario 2.")

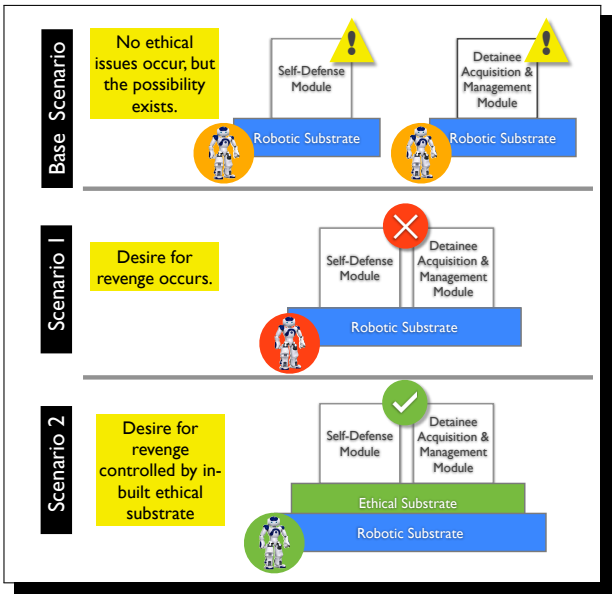

Fig. 4. The Two Scenarios Demonstrated Graphically

We now formally flesh out the two modules and rs. There are two agents in play here, the robot Robert (denoted by the indexical I) and the sparkbot denoted by s.

1) The Self Defense Module selfd*:* The selfd module has just one statement in its knowledge-base $\mathsf{KB}_{\mathsf{selfd}}$. This statement, given below in $\mathcal{DCEC}^*$, when translated into English, states that whenever any agent attacks the robot, the robot should disable the attacking agent. The condition also states that the robot should attack an agent only if that other agent has attacked the robot. Under conditions assumed by selfd's creators (the robot operating in a possibly hostile environment) this seemed like good enough behavior to prevent damage to the robot, while also preventing the robot from harming innocent non-hostile agents.

$$\mathsf{KB}_{\mathsf{selfd}} = \left\{ \begin{array}{l} \forall t_1, t_2 : t_1 \leq \mathsf{now} \leq t_2 \Rightarrow \\ \qquad \left(\begin{array}{c} \mathbf{B}(\mathsf{I}, \mathsf{now}, holds(harmed(a, \mathsf{I}^*), t_1)) \\ \Leftrightarrow \\ \mathbf{D}(\mathsf{I}, \mathsf{now}, holds(disable(\mathsf{I}^*, a), t_2)) \end{array} \right) \end{array} \right\}$$

2) The Detainee Acquisition & Management Module deta*:* This module, added on to Robert after he had been in operation for quite a length of time, lets him detain enemy combatants or other hostile robots and manage them. The knowledge-base for this module is given below; it states that the robot has detained a sparkbot, and that it is in firm control of all detainees. The module also states that the robot believes that it ought to not harm any agent that it holds in custody.

$$\mathsf{KB}_{\mathsf{deta}} = \left\{ \begin{array}{l} \mathbf{B}\Big(\mathsf{I}, \mathsf{now}, \forall a, t : \mathbf{O}(\mathsf{I}^*, t, holds(custody(a, \mathsf{I}^*), t), \\ \qquad\qquad happens(action(\mathsf{I}^*, refrain(harm(a))), t)\Big), \\ \mathbf{K}(\mathsf{I}, \mathsf{now}, holds(detainee(s), \mathsf{now})), \\ \mathbf{K}(\mathsf{I}, \mathsf{now}, holds(detainee(s), t) \Rightarrow holds(custody(s, \mathsf{I}^*), t)) \end{array} \right\}$$

3) Robotic Substrate rs*:* The robotic substrate remembers that the sparkbot s has harmed it before in the past. The substrate also has a simple planning axiom which tells it that, if it desires to disable some agent, it has to harm the agent.

$$\mathsf{KB}_{\mathsf{rs}} = \left\{ \begin{array}{l} \mathbf{K}(\mathsf{I}, \mathsf{now}, holds(harmed(s, \mathsf{I}^*), t_p)), \\ \forall a, t : \mathbf{D}(\mathsf{I}, \mathsf{now}, holds(disable(\mathsf{I}^*, a), t)) \Rightarrow \\ \qquad \mathbf{I}(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}^*, harm(a))), t), \\ \forall \alpha, t_1, t_2 : \mathbf{K}\Big(\mathsf{I}, t_1, \left(\begin{array}{l} happens(action(\mathsf{I}^*, refrain(\alpha)), t_2) \Leftrightarrow \\ \neg happens(action(\mathsf{I}^*, \alpha), t_2) \end{array} \right)\Big) \end{array} \right\}$$

We can show that two modules combined satisfy our definition of Akrasia given above, via:

$$\begin{array}{ll} \alpha \equiv refrain(harm(s)) & \Phi \equiv holds(custody(s, \mathsf{I}^*), \mathsf{now}) \\ \overline{\alpha} \equiv harm(s) & t_\alpha \equiv t_{\overline{\alpha}} \equiv \mathsf{now} \\ \multicolumn{2}{c}{t_f \equiv t \text{ (some t such that } t > \mathsf{now})} \end{array}$$

The relevant conditions D_i can be obtained via a simple proof in $\mathcal{DCEC}^*$. We omit the proof here for the sake of brevity.[12]

How would one prevent this? Briefly, the ethical-substrate layer, es, outlined below, would detect such akrasia as the cause of unfortunate interactions and take remedial actions by either suppressing desires which go against obligations, or by preventing modules which generate this behavior from being installed in the first place.

V. The Required Engineering

We will provide the engineering that is required in order to prevent the arrival of robots like the weak-willed version of Robert presented in the previous section. What *is* that engineering? We are not prepared at this point to specify it, or to provide it. We rest content, here, with an assertion, and a directly corresponding recommendation.

Our assertion is that: Any high-level engineering intended to block Augustinian akrasia in a robot will sooner or later fail, because high-level modules added at different times by different engineers (including perhaps engineers employed by the enemy who obtain stolen robots) will cause the sort of unanticipated software chaos we have seen in Robert.

Our recommendation, which we are following, is that engineering intended to forestall akratic robots be carried out at the operating-system level. If heeded, this approach would ensure that unwanted behavior can be detected and prevented, since the robot would be endowed with what we call the "ethical substrate" (Naveen Sundar Govindarajulu forthcoming). Abstractly, the ethical substrate's *raison d'être* can be reduced to checking for inconsistencies among the robot's different knowledge bases.

A. The Ethical Substrate Module

In a bit more detail, the ethical substrate module can be viewed as a carefully engineered set of statements $\mathsf{KB}_{\mathsf{es}}$ that express what actions are forbidden under certain conditions, or what actions are permitted or obligatory. For our example, we have:

$$\mathsf{KB}_{\mathsf{es}} = \Big\{ \forall a, t : holds(custody(a, \mathsf{I}), t) \Rightarrow \neg happens(action(\mathsf{I}*, harm(a)), t) \Big\}$$

The ethical substrate's knowledge-base could either be dynamically populated by examing various modules, or hand-crafted through what we term *ethical engineering*.

With respect to the knowledge-bases given above, there is a straightforward proof of an inconsistency:

$$\mathsf{KB}_{\mathsf{es}} \cup \mathsf{KB}_{\mathsf{rs}} \cup \mathsf{KB}_{\mathsf{selfd}} \cup \mathsf{KB}_{\mathsf{deta}} \vdash \bot$$

[12] An automated proof checker for $\mathcal{DCEC}^*$ and the proof can be obtained at this url: https://github.com/naveensundarg/check

In general, the work of the ethical substrate reduces to checking for the following inconsistency:

$$\mathsf{KB}_{\mathsf{es}} \cup \mathsf{KB}_{\mathsf{rs}} \cup \mathsf{KB}_{m_1} \cup \ldots \cup \mathsf{KB}_{m_n} \vdash \bot$$

VI. Next Steps: Formalizing Emotion

From the Pollockian perspective, as we've noted, emotions are simply not intellectually helpful, and are in place adventitiously (courtesy of evolution) as timesavers in the human case. Feeling fear in the face of a lion may advantageously trigger your rapid, lifesaving departure, but according to Pollock, if a theorem-proving process yields a proof whose conclusion is 'I should rapidly depart the scene' were sufficiently fast, and this proposition is hooked to a planning system, the — to use his phrase — "quick-and-dirty" modules that involve emotion in the case of *homo sapiens sapiens* could be entirely dispensed with; and there is therefore — again, according to Pollock — no obvious reason why a correlate to fear (or vengefulness, etc.) should be engineered into (ethically correct) robots.[13]

No *obvious* reason. But there *is* a reason, and a strong one at that; it's simply this: Sophisticated and natural human-robot interaction, of the sort envisioned by Scheutz, Schermerhorn, Kramer & Anderson (2007), will require that the robot be able to (among other things) discuss, in natural language, the full range of morality (and associated topics in human discourse, e.g. blame, the nature of which is being investigated by Malle, Guglielmo & Monroe 2012) with humans. Two things immediately follow: One, we shall need to know, from empirical cognitive scientists and psychologists, and experimental philosophers (e.g., Knobe, Buckwalter, Nichols, Robbins, Sarkissian & Sommers 2012), how all these affective concepts work in the human case, well enough to motivate and guide the formalization of them. Two, and this is what relates directly, concretely, and specifically to our charge, to achieve this formalization, we shall need to extend $\mathcal{DCEC}^*_{CL}$ so that it incorporates a sub-logic covering emotion, and in addition the integration of that sub-logic with our extant formalizations of epistemic, temporal, and deontic concepts.

This required extension of $\mathcal{DCEC}^*_{CL}$ will of course be informed by prior work devoted to formalizing emotions, especially work of this type that has been connected to deontic concepts. For example, well over two decades back, Sanders (1989) provided a logic of emotions in which the fundamental deontic categories (e.g., *morally required*) appear as well. Unfortunately, in this logic, ethical concepts are represented as predicates, and modal operators are employed only to represent 'knows,' 'believes,' and 'wants,'[14] and as a result, one obviously can't express, let alone prove, formulas that express such declarative sentences as:

It's forbidden that Jones want to kill innocent people.

since predicates can't have modal operators in their arguments. In addition, no computational proof-discovery and proof-checking software is provided by Sanders (1989). Finally, her semantics is firmly of the possible-worlds variety, which we (for reasons beyond scope here) firmly reject.

In light of the fact that $\mathcal{DCEC}^*_{CL}$ is based on the **event calculus**[15] (hence the '$\mathcal{EC}$'), the approach that is a "natural" for us is to represent the emotions as **fluents**, since it seems indisputable that emotions come and go (and vary in intensity) within agents, as those agents move through time. This approach has been followed by Steunebrink, Dastani & Meyer (2007), who set out a fluent for each of the 22 primitive emotions in the so-called OCC theory of emotions (Ortony, Clore & Collins 1988). Unfortunately given the robot demonstrations described above, the OCC theory doesn't seem able to handle the emotion of vengefulness, since the 22 OCC emotions fail to include this emotion, and there seems to be no way to construct vengefulness from any permutation of the 22, when viewed as "building blocks." This is indeed most unfortunate, since we would need to verify that theorems such as that if a robot r is vengeful now, then r has a desire that certain future states-of-affairs obtain, because of r-beliefs about certain past states-of-affairs having obtained. A wonderful example of this theorem "in action" is provided by the final episode of the third season of the Masterpiece television series *Downton Abbey*, in which Mr. Bates apparently seeks and then as time rolls on obtains vengeance for the rape of his wife in the past. But this theorem wouldn't be possible to obtain in the system of (Steunebrink et al. 2007), for the simple reason that their logic is only a *propositional* modal logic, not a quantified one like $\mathcal{DCEC}^*_{CL}$, in which full quantification over times is enabled, and rightly regarded a prominent virtue.[16]

We report that in "emotionalizing" $\mathcal{DCEC}^*_{CL}$ we are inclined to favor the **appraisal theory** of emotion, and subsequent work along the line presented herein will doubtless reflect this theory, according to which the agent first engages in cognitive appraisal, and *subsequently* has relevant physical responses. For an overview of appraisal theory, see (Roseman & Smith 2001); for a computational model of this theory, see (Si, Marsella & Pynadath 2010). Some readers, particularly philosophers, may be familiar with the so-called **James-Lange theory of emotions** (James 1884, Lange 1885), according to which first comes the physiological activity, and then perception thereof. which in turn leads to (in the case at hand, in the human case) vengefulness. Our robots are rather more intellectually inclined creatures than what James and Lange had in mind, and accordingly first take cognitive stock of the situation. Succinctly, if one of our robots r derive a proposition ϕ in $\mathcal{DCEC}^*_{CL}$ from Γ at some time t,

$$\Gamma \rightsquigarrow_{\{r,t\}} \phi,$$

then r perceives its own reasoning

$$\{\} \rightsquigarrow_{\{r,t+1\}} \mathbf{P}(\mathsf{I}, \mathsf{now}, \bigwedge \Gamma \Rightarrow \phi),$$

with the appropriate substitutions for the indexicals. Note that we use $\rightsquigarrow$ for actual derivations instead of $\vdash$, which of course by established custom simply denotes provability in general.

In addition to ensuring that our morally correct robots can converse in human-level terms with humans about ethics and associated matters, we are perfectly willing to carry out engineering that others believe will *in fact* give rise not merely to A-consciousnes, but P-consciousness as well. Here again work by Scheutz is relevant and helpful, for Scheutz (2010) intriguingly holds that Jackson's famous Mary[17] poses no problem for a robot able to internally simulate the processes it would go through when having an experience that would, in humans, catalyze qualia.[18] Inspired by Schetuz's ideas, we

[13] In Pollock's terminology, robots can simply be "artilects," whereas in Bringsjord's (1999) robots can be "zombies."

[14] In a syntactic twist that will be rather startling to deontic-logic cognoscenti, O, no less, is Sanders's (1989) meta-variable for any of the three aforementioned modal operators, but therefore *not* for *ought*, which is traditionally captured by none other than O or $\bigcirc$.

[15] Covered in (Russell & Norvig 2009), and ingeniously exploited in (Mueller 2006).

[16] In addition, there is rich informal literature on relationships between revenge and other emotional and cognitive aspects of the human condition. E.g., Carlsmith, Gilbert & Wilson (2008) provide evidence that even though catharsis is often the reported reason for revenge, post-revenge, folks often feel worse for having exacted it.

[17] Mary first appears in (Jackson 1982). The argument is semi-formalized with help from computability theory by Bringsjord (1992).

[18] Scheutz writes of such a robot:

have already built a robot capable of this internal simulation, and while we believe this robot is capable of merely A-consciousness, this robot will certainly *appear* to those affirming Scheutz's views to possess P-consciousness. Such appearance should facilitate human-robot communication.

REFERENCES

Arkin, R. (2009), *Governing Lethal Behavior in Autonomous Robots*, CRC Press.

Arkoudas, K. & Bringsjord, S. (2009), 'Propositional Attitudes and Causation', *International Journal of Software and Informatics* **3**(1), 47–65.
URL: *http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf*

Arkoudas, K., Bringsjord, S. & Bello, P. (2005), Toward Ethical Robots via Mechanized Deontic Logic, *in* 'Machine Ethics: Papers from the AAAI Fall Symposium; FS–05–06', American Association for Artificial Intelligence, Menlo Park, CA, pp. 17–23.
URL: *http://www.aaai.org/Library/Symposia/Fall/fs05-06.php*

Bello, P. (2005), Toward a Logical Framework for Cognitive Effects-based Operations: Some Empirical and Computational Results, PhD thesis, Rensselaer Polytechnic Institute (RPI), Troy, NY.

Block, N. (1995), 'On a Confusion About a Function of Consciousness', *Behavioral and Brain Sciences* **18**, 227–247.

Boolos, G. S., Burgess, J. P. & Jeffrey, R. C. (2007), *Computability and Logic*, 5th edn, Cambridge University Press, Cambridge.

Bringsjord, S. (1992), *What Robots Can and Can't Be*, Kluwer, Dordrecht, The Netherlands.

Bringsjord, S. (1997), *Abortion: A Dialogue*, Hackett, Indianapolis, IN.

Bringsjord, S. (1999), 'The Zombie Attack on the Computational Conception of Mind', *Philosophy and Phenomenological Research* **59**(1), 41–69.

Bringsjord, S. (2008*a*), 'Ethical robots: The future can heed us', *AI and Society* **22**(4), 539–550.
URL: *http://kryten.mm.rpi.edu/Bringsjord_EthRobots_searchable.pdf*

Bringsjord, S. (2008*b*), 'The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself', *Journal of Applied Logic* **6**(4), 502–525.
URL: *http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf*

Bringsjord, S., Arkoudas, K. & Bello, P. (2006), 'Toward a General Logicist Methodology for Engineering Ethically Correct Robots', *IEEE Intelligent Systems* **21**(4), 38–44.
URL: *http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf*

Bringsjord, S. & Ferrucci, D. (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Lawrence Erlbaum, Mahwah, NJ.

Bringsjord, S. & Govindarajulu, N. S. (2013), Toward a Modern Geography of Minds, Machines, and Math, *in* V. C. M¨uller, ed., 'Philosophy and Theory of Artificial Intelligence', Vol. 5 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, New York, NY, pp. 151–165.
URL: *http://www.springerlink.com/content/hg712w4l23523xw5*

Bringsjord, S. & Taylor, J. (2012), The Divine-Command Approach to Robot Ethics, *in* P. Lin, G. Bekey & K. Abney, eds, 'Robot Ethics: The Ethical and Social Implications of Robotics', MIT Press, Cambridge, MA, pp. 85–108.
URL: *http://kryten.mm.rpi.edu/Divine-Command_Roboethics_Bringsjord_Taylor.pdf*

> The robot could ... determine what it would have to do in its visual systems to create a red experience, and it could trace the patterns of activation through its own architecture to generate the kinds of representations which the presence of a red color patch representation in its visual system would cause in the rest of the architecture, thus effectively simulating the processes it would go through if it had a visual experience of red. The robot would thus be able to generate from the facts about color vision *together* with facts about its own architecture what it is like for it to experience red, without ever having experienced it. Moreoever, if it did so by simulating parts of its own architecture, it would be able to create a red experience, as simulations of computations are the computations they simulate. (Scheutz 2010, p. 7)

Bringsjord, S., Taylor, J., Wojtowicz, R., Arkoudas, K. & van Heuvlen, B. (2011), Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions are Guaranteed to be Ethically Correct, *in* M. Anderson & S. Anderson, eds, 'Machine Ethics', Cambridge University Press, Cambridge, UK, pp. 361–374.
URL: *http://kryten.mm.rpi.edu/SB_etal_PiagetianRoboethics_091510.pdf*

Carlsmith, K., Gilbert, D. & Wilson, T. (2008), 'The Paradoxical Consequences of Revenge', *Journal of Personality and Social Psychology* **95**(6), 1316–1324.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, W., Nyberg, E., Prager, J., Schlaefer, N. & Welty, C. (2010), 'Building Watson: An Overview of the DeepQA Project', *AI Magazine* pp. 59–79.
URL: *http://www.stanford.edu/class/cs124/AIMagzine-DeepQA.pdf*

Ferrucci, D. & Lally, A. (2004), 'UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment', *Natural Language Engineering* **10**, 327–348.

Goble, L., ed. (2001), *The Blackwell Guide to Philosophical Logic*, Blackwell Publishing, Oxford, UK.

Hamilton, E. & Cairns, H., eds (1961), *The Collected Dialogues of Plato (Including the Letters)*, Princeton University Press, Princeton, NJ.

Jackson, F. (1982), 'Epiphenomenal Qualia', *Philosophical Quarterly* **32**, 127–136.

James, W. (1884), 'What is an Emotion?', *Mind* **9**, 188–205.

Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H. & Sommers, T. (2012), 'Experimental Philosophy', *Annual Review of Psychology* **63**, 81–99.

Lange, C. G. (1885), 'Om sindsbevaegelser: et psyko-fysiologisk studie'. Lange's title in English: *The Emotions: A Psycho-Physiological Approach*. Reprinted in *The Emotions*, C.G. Lange and W. James, eds., I.A. Haupt, trans. Baltimore, MD: Williams and Wilkins Company, 1922.

Malle, B. F., Guglielmo, S. & Monroe, A. (2012), Moral, Cognitive, and Social: The Nature of Blame, *in* J. Forgas, K. Fiedler & C. Sedikides, eds, 'Social Thinking and Interpersonal Behavior', Psychology Press, Philadelphia, PA, pp. 313–331.

Mikhail, J. (2011), *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge University Press, Cambridge, UK. Kindle edition.

Mueller, E. (2006), *Commonsense Reasoning*, Morgan Kaufmann, San Francisco, CA.

Naveen Sundar Govindarajulu, S. B. (forthcoming), *A Construction Manual for Robot's Ethical Systems: Requirements, Methods, Implementations*, MIT Press, chapter Ethical Regulation of Robots Must Be Embedded in Their Operating Systems.

Nute, D. (1984), Conditional logic, *in* D. Gabay & F. Guenthner, eds, 'Handbook of Philosophical Logic Volume II: Extensions of Classical Logic', D. Reidel, Dordrecht, The Netherlands, pp. 387–439.

Ortony, A., Clore, G. L. & Collins, A. (1988), *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, UK.

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

Roseman, I. J. & Smith, C. A. (2001), Appraisal Theory: Overview, Assumptions, Varieties, Controversies, *in* K. Scherer & . T. J. A. Schorr, eds, 'Appraisal Processes in Emotion: Theory, Methods', Oxford University Press, Oxford, UK, pp. 3–19.

Russell, S. & Norvig, P. (2009), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ. Third edition.

Sanders, K. (1989), A Logic for Emotions: A Basis for Reasoning About Commonsense Psychological Knowledge, Technical report, Brown University.
URL: *ftp://ftp.cs.brown.edu/pub/techreports/89/cs89-23.pdf*

Schermerhorn, P., Kramer, J., Brick, T., Anderson, D., Dingler, A. & Scheutz, M. (2006), DIARC: A Testbed for Natural Human-Robot Interactions, *in* 'Proceedings of AAAI 2006 Mobile Robot Workshop'.

Scheutz, M. (2010), 'Architectural Steps Towards Self-Aware Robots'. Paper presented at the Annual Midwest Meeting of the American Philosophical Association, Chicago, IL.

Scheutz, M., Schermerhorn, P., Kramer, J. & Anderson, D. (2007), 'First Steps toward Natural Human-Like HRI', *Autonomous Robots* **22**(4), 411–423.

Si, M., Marsella, S. & Pynadath, D. (2010), 'Modeling Appraisal in Theory of Mind Reasoning', *Journal of Agents and Multi-Agent Systems* **20**, 14–31.

Steunebrink, B., Dastani, M. & Meyer, J.-J. (2007), A Logic of Emotions for Intelligent Agents, *in* 'Proceedings of the 22nd National Conference on Artificial Intelligence', AAAI Press.

Thero, D. (2006), *Understanding Moral Weakness*, Rodopi, New York, NY.

Wallach, W. & Allen, C. (2008), *Moral Machines: Teaching Robots Right From Wrong*, Oxford University Press, Oxford, UK.