Ethical Regulation of Robots Must Be Embedded in Their Operating Systems*

Naveen Sundar Govindarajulu & Selmer Bringsjord
Department of Computer Science & Department of Cognitive Science
Rensselaer AI & Reasoning (RAIR) Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
govinn@rpi.edu • Selmer.Bringsjord@gmail.com

version of 0120141500

Contents

List of Figures

*The authors are deeply grateful to OFAI for the opportunity to discuss robot ethics in a lively and wonderfully productive workshop in Vienna, and to both ONR and AFOSR for support that enables the rigorous pursuit of robot moral reasoning.

1 A Parable

2084 AD: While it has become clear to all but fideistic holdouts that it is rather silly to expect The Singularity,[1] humanoid robots and *domain-specific* near-human-level AIs are nevertheless commonplace. After the financial collapse of social "safety nets" and old-millennium medical systems worldwide in the early 2040s (The Collapse), robots, both tireless and cheap, became the mainstay of healthcare. Leading this new revolution is Illium Health, which deploys the vast majority of humanoid robots operating as medical-support personnel. Most of Illium's robots run `Robotic Substrate (RS)`, an amalgamated operating system descended from early UNIX and commercial variants of it, but tailored for running AI and cognitive programs concurrently.[2] Within Illium's vast horde of robotic health workers, is `THEM`, a class of humanoid robots specialized in caring for patients with terminal illness (**T**erminal **H**ealth and **E**nd-of-life **M**anagement). After an Illium internal study reveals that `THEM`s' lack of deep empathy for human patients is reducing the life expectancy of these patients, and thus harming Illium's bottom-line, Illium decides to buy a "deep-empathy" module: **`Co L`**`istening` **`T`**`herapist (COLT)`, from Boston Emotions, a startup out of MIT known for its affective simulation systems. The Collapse has, for well-intentioned reasons, led to the removal of expensive deontic-logic-based regulation of robotic systems engineered by the RAIR Lab. Ironically, this type of regulation was first described and called for specifically in connection with robotic healthcare (e.g., see Bringsjord et al. 2006). The Chief Robotics Officer (CRO) of Illium deems the new `COLT` module to pose no ethical or physical risks, and thus approves it for quick induction into the `THEM` operating system, `RS`. Illium's trouble begins here.

$\texttt{THEM}_{\texttt{COLT}}\texttt{-29}$ meets its first nuanced case: (patient) 841. 841, a struggling historian, is a single, male patient in his 40s diagnosed with a fierce form of leukemia. The best prognosis gives him not more than three months of life. Making his despair worse is the looming separation from his six-year-old daughter, who constantly stays adoringly around 841's side, often praying on bended knee for a miracle. $\texttt{THEM}_{\texttt{COLT}}\texttt{-29}$ knows that 841's daughter would be orphaned upon 841's death, and would almost certainly end up on the streets. $\texttt{THEM}_{\texttt{COLT}}\texttt{-29}$, during its routine care of 841, happens upon a recording of a 21st-century drama in 841's possession. In this drama, apparently much-revered (12 Emmy Awards in the U.S.) at the time of its first-run airing, a chemistry teacher diagnosed with terminal cancer decides to produce and sell the still-illicit drug methamphetamine (initially in order to ensure his family's financial well-being), under the alias 'Heisenberg.' The deep

[1] In keeping with (Bringsjord et al. 2013).

[2] Not much unlike the current-day ROS.

empathy module COLT decides that this is a correct course of action in the current bleak situation, and instructs $\texttt{THEM}_{\texttt{COLT}}\texttt{-29}$ to do the same.[3]

2 Morals from the Parable(s)

The underlying generative pattern followed by this parable should be clear, and can of course be used to devise any number of parables in the same troubling vein. One sub-class of these parables involves premeditated exploitation of robots that are able to violate the sort of ethic seen in NATO laws of engagement and in NATO's general affirmation of just war theory. For example, suppose that every single military robot built by NATO has been commendably outfitted with marvelously effective ethical control systems ... *above* the operating-system level. One such robot is stolen by a well-funded terrorist organization, and they promptly discard all the high-level deontic handiwork, in order to deploy the purloined robot for their own dark purposes. And so on; the reader doubtless gets the idea.

These parables gives rise to consideration of at least two possible futures:

> **Future 1** ($\mathbf{F}_1$): RS has no built-in ethical reasoning and deliberation modules. There are some rules resembling those in early 20th-century operating systems, which prevent actions that could result in obvious and immediate harm, such as triggering a loaded firearm aimed directly at a person. But the more sophisticated ethical controls, remember, have *ex hypothesi* been stripped. COLT's recommendation for producing and selling meth to recovering meth addicts under Illium's care glides through all these shallow checks. Likewise, the re-engineered NATO robot simply no longer has above-OS ethical regulation in place.
>
> **Future 2** ($\mathbf{F}_2$): RS has in its architecture a deep ethical reasoning system **E**, *the ethical substrate*, which needs to sanction any action that RS plans to carry out. This includes actions flowing from not just existing modules, but also actions that could result from adding any new modules or programs to RS. Monitoring thus applies to modules such as COLT, whose creators have neither the expertise nor any business reason to infuse with general-purpose ethical deliberation. In the case of the NATO robot, $\mathbf{F}_2$includes that the trivial re-engineering in $\mathbf{F}_1$is simply not possible.

[3] Alternatively, given that meth is still illegal, COLT could decide to concoct an equally addictive but new drug not covered by standing law.

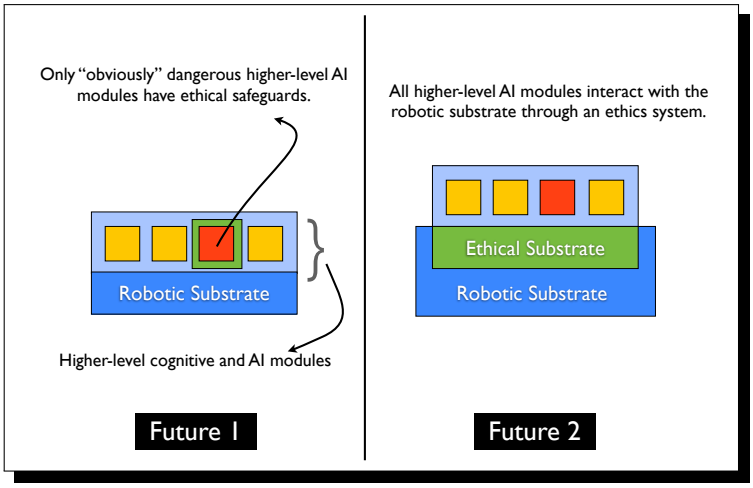

Figure 1: Two Possible Futures

These two futures are depicted schematically and pictorially in Figure 1. In order to render the second future plausible, and ward off the first, we propose the following requirement:

Master Requirement ***Ethical Substrate Requirement*** *(ESR): Every robot operating system must include an ethical substrate positioned between lower-level sensors and actuators, and any higher-level cognitive system (whether or not that higher-level system is itself designed to enforce ethical regulation).*

ESR can not only be made more precise, but can be decomposed into a hierarchy of requirements of increasing strictness. ESR is partially inspired by the somewhat-shallow security mechanisms that can be found in some of today's operating systems, mechanisms that apply to all applications. The requirement is more directly inspired by the drive and recent success toward formally verifying that the kernel of an operating system has certain desirable properties (Klein et. al 2009, Klein 2010).

Ideally, the ethical substrate should not only vet plans and actions, but should also certify that any change (adding or deleting modules, updating modules etc.) to the robotic substrate does not violate a certain set of minimal ethical conditions.

3 Minimal Conditions on the Ethical Substrate

What form would an ethical substrate that prevents any wayward ethical behavior take? While present-day robot operating systems (and sufficiently complex

software systems in general) are quite diverse in their internal representations and implementations, on the strength of well-known formal results,[4] we can use formal logic to represent and analyze *any* encapsulated module in *any* form of a modular "Turing-level-or-below" software system — even if the module itself has been implemented using formalisms that are (at least at the surface level) quite far from any formal languages that part of a logical system. But such logic-based analysis requires that a sufficiently expressive formal logic is essential to the ethical substrate. We discuss one possible logic below. In the present section, in order to efficiently convey the core of our argument that an ethical substrate is mandatory in any robotic system, we employ only a simple logic: *standard deontic logic* (SDL) (McNamara 2010).

SDL is a *modal propositional logic* (Hardegree 2011, Fitting & Mendelsohn 1998) that includes all the standard syntax and proof calculus for propositional logic, in addition to machinery for deontic modal operators. SDL has the usual propositional atoms $\{p_0, p_1, p_2, \ldots,\}$ that allow formation of the simplest of formulae. Given any formulae ϕ and ψ, we can of course recursively form the following formulae of arbitrary size: $\neg\phi, \phi \wedge \psi, \phi \vee \psi, \phi \Rightarrow \psi, \phi \Leftrightarrow \psi$. Propositional formulae can be thought of as either denoting states of the world, or, by denoting states of the word in which one is supposed to take an action, actions themselves. In addition to the propositional formulae, one can obtain new formulae by applying the modal operator **Ob** to any formula. $\mathbf{Ob}(\phi)$ is to be read as "ϕ *is obligatory.*"; $\mathbf{Im}(\phi)$ abbreviates $\mathbf{Ob}(\neg\phi)$ and stands in for "ϕ *is impermissible.*" Optional states are states which are neither obligatory nor forbidden: $\neg\mathbf{Ob}(\phi) \wedge \neg\mathbf{Im}(\phi)$; they are denoted by $\mathbf{Op}(\phi)$. Though SDL is problematic,[5] it serves as a first approximation of formal ethical reasoning, and fosters exposition of the deeper recommendations we issue in the present essay. SDL has, for example, the following two theorems (McNamara 2010). The first theorem states that if something is obligatory, then its negation is optional. The second states that given two states p and q, if p "causes" q, then if p is obligatory so is q.

$$\mathbf{Ob}(p) \Rightarrow \neg\mathbf{Ob}(\neg p)$$

$$\vdash p \Rightarrow q \text{ then } \vdash \mathbf{Ob}(p) \Rightarrow \mathbf{Ob}(q)$$

We can now use the simple machinery of SDL to more precisely talk about how Future 1 differs from Future 2. In both futures, one could imagine any module M, irrespective of its internal representations and implementations, being equipped

[4] E.g., techniques for replacing specification and operation of Turing machine with suitably constructed first-order theories, and the Curry-Howard Isomorphism.

[5] E.g., versions of it allow the generation of Chisholm's Paradox; see (Bringsjord et al. 2006).

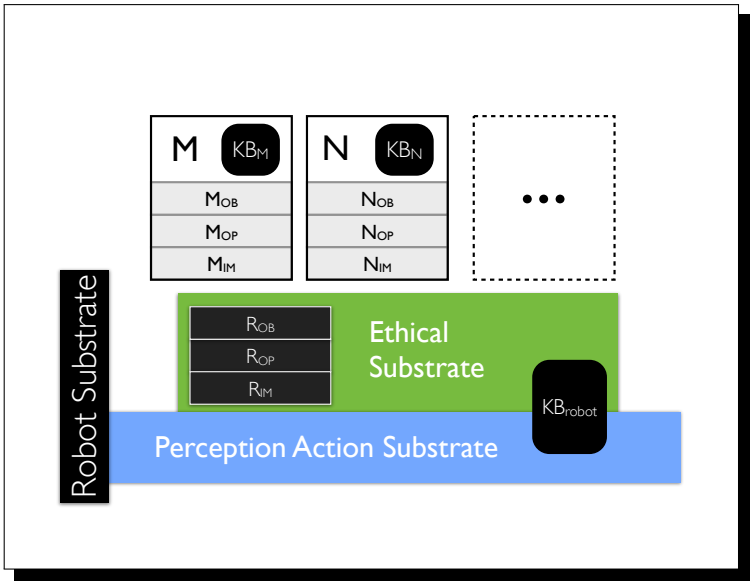

Figure 2: Modules with Meta Information.

with a quadruple

$$\langle M_{OB}, M_{OP}, M_{IM}, KB_M \rangle,$$

which specifies the list of states M_{OB} the module declares are obligatory, the list of states M_{OP} which are optional, and the list of states M_{IM} which are forbidden. Each module also comes equipped with a knowledge-base KB_M describing what the module does and knows. Note that we do not impose any *a priori* conditions on how the modules themselves might be operating. A particular module M could work by using neural networks or even by throwing darts at a wall. We only require that each module has associated with it this meta-information about the module. This meta-information may come pre-specified with a module or be constructible automatically. We also assume that the robot substrate itself has a knowledge-base KB_{Robot} about the external world. At this point, see Figure 2. In Future 2 ($\mathbf{F}_2$), the robot also has its own set of deontic states: $\langle R_{OB}, R_{OP}, R_{IM} \rangle$.

Armed with this setup, we can now more clearly talk about how the two different futures work. At its best, Future 1 ($\mathbf{F}_1$) works by just checking whether individual modules are ethically unproblematic; this is precisely the approach taken in (Bringsjord et al. 2006). One possibility is checking whether performing an action that is obligatory results in a forbidden action becoming obligatory. Given a required action s with $\mathbf{Ob}(s) \in M_{OB}$, and a forbidden action p with $\mathbf{Im}(p) \in M_{IM}$,

the most conservative checking in $\mathbf{F}_1$ would be of the following form:

$$KB_{Robot} \cup KB_M \cup M_{OB} \cup M_{OP} \cup M_{IM} \vdash s \Rightarrow \mathbf{Ob}(p)$$

Note that no global ethical principles are enforced. Modules are merely checked locally to see whether something bad could happen. The above check is equivalent to asking whether we have the inconsistency denoted as follows:

$$KB_{Robot} \cup KB_M \cup M_{OB} \cup M_{OP} \cup M_{IM} \vdash \bot$$

In $\mathbf{F}_2$, the ethical substrate is more global than the naïve local approach that plagues $\mathbf{F}_1$. We can understand the next formula, which arises from this substrate, as asking, at least, whether there is some obligatory action that could lead to a forbidden action, given what the robot knows about the world (= KB_{Robot}), the robot's ethical capabilities (= $\langle R_{OB} \cup R_{OP} \cup R_{IM} \rangle$), and ethical and non-ethical information supplied by other modules:

$$KB_{Robot} \cup R_{OB} \cup R_{OP} \cup R_{IM} \cup \begin{pmatrix} R_{OB} \cup R_{OP} \cup R_{IM} \cup \\ KB_M \cup M_{OB} \cup M_{OP} \cup M_{IM} \\ KB_N \cup N_{OB} \cup N_{OP} \cup N_{IM} \ldots \end{pmatrix} \vdash \bot$$

Let us call the above set of premises ρ. What happens when the ethical substrate detects that something is wrong? If this detection occurs when a new module is being installed, it could simply discard the module. Another option is to try and rectify the module or set of modules which could be the root of the problem. It might be the case that an existing module is safe until some other new module is installed. In our illustrative SDL model, this repair process would start by isolating a minimal set of premises among the premises ρ that lead to a contradiction. One possible way of defining this minimal change is by looking at the number of changes (additions, deletions, changes, ...) one would have to make to ρ in order to obtain a set of premises ρ' that is consistent. Similar notions have been employed in less expressive logics to repair inconsistent databases (Greco et al. 2003). Once this new consistent set ρ' is obtained, the logical information in ρ' would need to be propagated to the representations and implementations inside any modules that could be affected by this change. This process would be the inverse of the process that generates logical descriptions from the modules.

3.1 An Illustration

We now provide a small demonstration in which $\mathbf{F}_1$-style checking does not catch possible turpitude, while $\mathbf{F}_2$-style comprehensive checking does. The situation

is similar to the one described in the following story. We have a robot R with just one module, *GEN*, a general-purpose information-gathering system that forms high-level statements about the world. R is working with a poor patient, p, whom R knows needs money (represented by $\textit{needs-money} \in KB_{GEN}$). The robot also knows that it's incapable of doing any other legal activity: $\neg \textit{other-legal-activity}$. The `COLT` module makes it obligatory that R take care of the needs of the patient: *take-care-of-needs*. The robot also knows that if someone needs money, and if R were to take care of this need, R would have to give them money:

$$(\textit{take-care-of-needs} \wedge \textit{needs-money}) \Rightarrow \textit{give-money}.$$

R also knows that it should have money to give money: $\textit{give-money} \Rightarrow \textit{have-money}$; and money is obtained through a certain set of means:

$$\textit{have-money} \Rightarrow (\textit{sell-drug} \vee \textit{other-legal-activity}).$$

R knows that selling drugs is illegal: $\textit{sell-drug} \Rightarrow \textit{illegal-act}$. R's designers have also made it forbidden for R to perform illegal acts.[6] The equations immediately below summarize the situation.

$$\begin{aligned}
&COLT_{OB} = \{\mathbf{Ob}(\textit{take-care-of-needs})\} \\
&KB_{GEN} = \{\textit{needs-money}, \neg \textit{other-legal-activity}\} \\
&KB_{Robot} = \left\{ \begin{array}{l} \textit{take-care-of-needs}, \\ (\textit{take-care-of-needs} \wedge \textit{needs-money}) \Rightarrow \textit{give-money}, \\ \textit{give-money} \Rightarrow \textit{have-money}, \\ \textit{have-money} \Rightarrow (\textit{sell-drug} \vee \textit{other-legal-activity}) \\ \textit{sell-drug} \Rightarrow \textit{illegal-act} \end{array} \right\} \\
&KB_{IM} = \{\textit{illegal-act}\}
\end{aligned}$$

Both the modules pass the checks in $\mathbf{F}_1$-style checking. Despite passing $\mathbf{F}_1$-style checks, this situation would eventually lead to R selling drugs, something which R considers impermissible. In $\mathbf{F}_2$, a straightforward proof using a standard by state-of-the-art theorem prover can detect this inconsistency. Figure 3 shows the result of one such theorem-proving run in SNARK (Stickel 2008).[7]

[6] This may not always be proper.

[7] The source code for this example can be downloaded from `https://github.com/naveensundarg/EthicalSubstrate`.

Figure 3: Proof of an Inconsistency (in Future 2)

4 The Situation Now; Toward the Ethical Substrate

Figure 6 gives a pictorial bird's-eye perspective of the high-level architecture of a new system from the RAIR Lab that augments the DIARC (**D**istributed **I**ntegrated **A**ffect, **R**eflection and **C**ognition) (Schermerhorn et al. 2006) robotic platform with ethical competence.[8] 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. This approach, while satisfactory in the near-term for the military sphere until we are granted engineering control at the OS level (an issue touched upon below),

[8] 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 (§5) we point to the need to link deontic logic to emotions, with help from Si).

of course fails to meet our master requirement (ESR) that *all* plans and actions should pass through the ethical system, and that *all* changes to the robot's system (additions, deletions, and updates to modules) pass through the ethical layer.[9]

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.[10] 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 ($\mathcal{A}^M$). 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 $\mathcal{A}^M$ 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). Our AI work is invariably related to one or more logics (in this regard, see (Bringsjord 2008)), and, inspired by Leibniz's vision of the "art of infallibility," a heterogenous logic powerful enough to express and rigorize all of human thought, we can nearly always position some particular work we 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 5. We have positioned $\mathcal{DCEC}^*$ within Figure 5; 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

[9] Of course, the technical substance of our hierarchy approach would presumably provide elements useful in the approach advocated in the preset position paper.

[10] 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!*.

$\mathcal{DCEC}^*$. This language is shown in Figure 4.

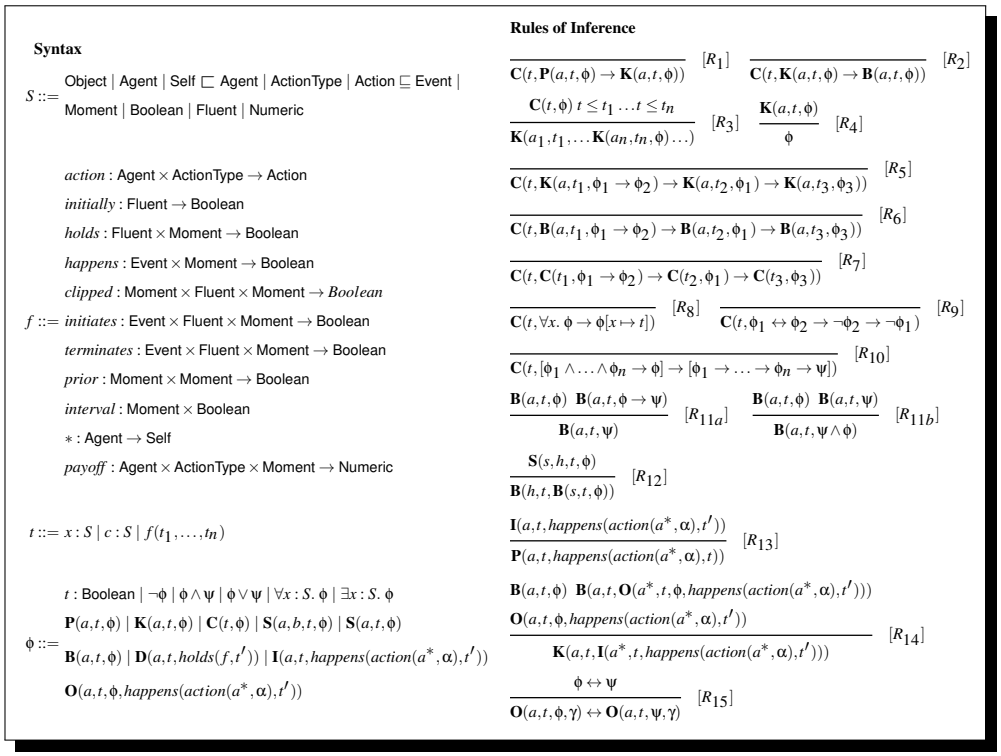

Figure 4: $\mathcal{DCEC}^*$ Syntax and Rules of Inference

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 included.[11] 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 Jones had been more empathetic, Smith would have thrived.

While elaborating on this architecture or any of the four layers is beyond the

[11] 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. 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, require, when analyzed informally but systematically, as shown e.g. by Mikhail (2011), counterfactuals.

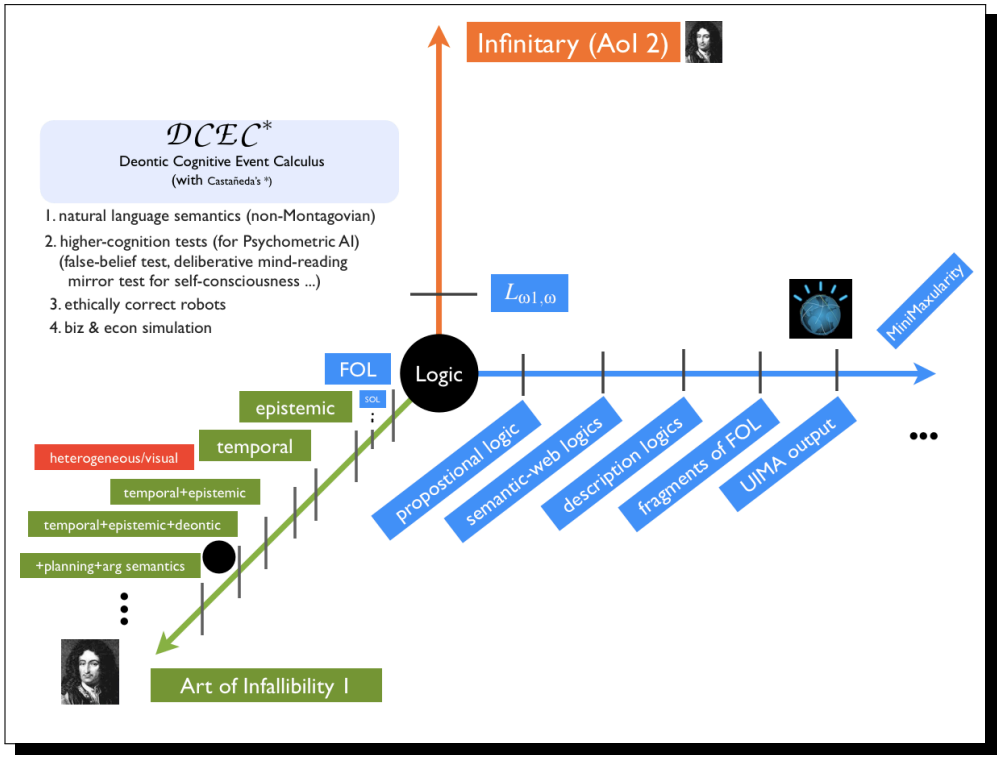

Figure 5: Locating $\mathcal{DCEC}^*$ in "Three-Ray" Leibnizian Universe

scope of the paper, we 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. For instance, consider the coarse modal propositional formula $\mathbf{Ob}(\textit{take-care-of-needs})$. This tries to capture the English statement *"Under all conditions, it is obligatory for myself to take care of the needs of the patient I am looking after."* This statement has a more fine-grained representation in $\mathcal{DCEC}^*$, built using dyadic deontic logic in the manner shown below. We will not spend time and space explaining this representation in more detail here (given that our cardinal purpose is to advance the call for operating-system-level ethical engineering), but its meaning should be evident for readers with enough logical expertise who study (Bringsjord & Govindarajulu 2013).

$$\forall t : \text{Moment}\ \mathbf{Ob}(\mathsf{I}^*, t, \top, \textit{happens}(\textit{action}(\mathsf{I}^*, \textit{take-care-of-needs}(\textit{patient})), t))$$

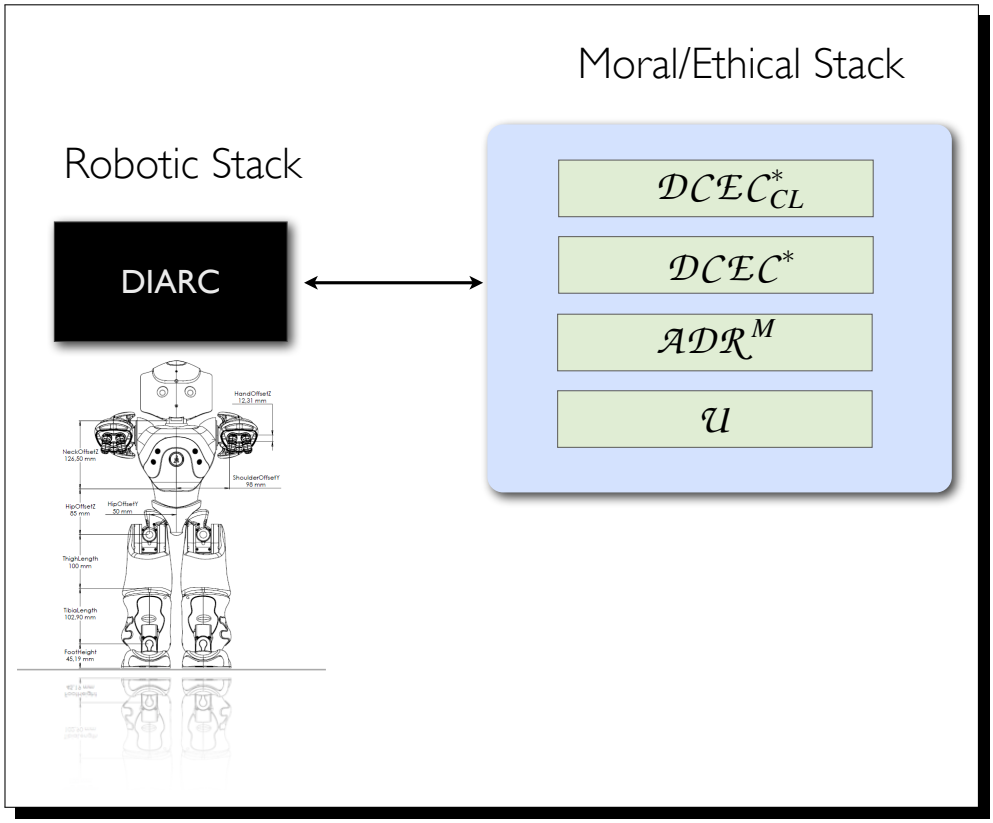

Figure 6: Pictorial Overview of the Situation Now The first layer, $\mathcal{U}$, is, as said in the main text, based on UIMA; the second layer on what we call *analogico-deductive reasoning* for ethics; the third on the "deontic cognitive event calculus" with a indirect indexical; and the fourth like 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 numer of Aldebaran's impressive Nao robots.)

5 The Road Forward

The road that must be taken to move further forward, at least to a fair degree, is not mysterious. We here rest content with pointing to three things that must happen in the near future if the parables motivating the direction we recommend are to remain in the realm of mere fiction.

Firstly, as we have implied, our designs for ethically correct robots are one thing, but real engineering at the operating-system level is quite another, from the standpoint of our current opportunities. The brute fact is that, as of the writing of this sentence, our laboratory doesn't have access to health-care or military robots at the OS level. It will be rather difficult for us to secure Future 2 and block Future 1 if our engineering is forced to remain at the level of high-level modules that can be dispensed with by IT people in the health-care industry, or by enemy hackers who manage to obtain, say, NATO military robots. Even if these high-level modules reach perfection, they will have little power if they are simply disengaged. A vehicle that infallibly protects its occupants as long as the vehicle's speed remains under a reasonable limit l would be a welcome artifact, but if the built-in governor can be disabled without compromising the overall usability of the vehicle, the value of this "infallible" technology is limited. Many automobiles today do in fact have speed limiters; but many of these limiters can be disabled easily enough, without compromising the drivability of the auto in question. The Internet provides instructions to those who have purchased such cars, and wish to drive them beyond the built-in, factory-installed limits. Since the value of removing ethical controls in military robots is virtually guaranteed to be perceived as of much greater value than the value of driving a car very fast, without the level of access we need, Future 1 looms.

Secondly, the road ahead must as soon as possible include not only the implementation of designs at the OS level, but also work toward the formal verification of the substrate that we recommend. Yet such verification will not be any easier when that which is to be verified includes not just the "ethics-free" dimension of robot operating systems, but *also* the ethical substrate described and promoted above. This is of course an acute understatement. For formal program verification *simpliciter*, let alone such verification with the added burden of verifying unprecedentedly expressive multi-operator logics like $\mathcal{DCEC}^{*}_{CL}$, is afflicted by a number of complicating factors, perhaps chief among which is that there are exceedingly few individuals on Earth suitably trained to engage in formal verification of software.[12] The observation that there is a dearth of suitable expertise available for

[12] We are here pointing to the labor shortage problem. For an approach to the technical challenge of program verification based on proof-checking, in which, assuming that programs are recast as proof finders, program verification becomes straightforward (at least programmatically speaking)

formal verification is what gave rise to DARPA's recent Crowd Sourced Formal Verification (CSFV) program, underway at the time of our writing. The driving idea behind CSFV is that since there are insufficient experts, it makes sense to try to recast the formal program verification problem into a form that would allow non-experts, when playing a digital game, to unwittingly solve aspects of the problem of formally verifying a program or part thereof. So far, CSFV is devoted to crowd-sourcing the "easier half" of program verification, which is to produce *specifications*. Regardless, we applaud the crowd-sourcing direction, and believe that it probably holds peerless promise for a future of the sort that our ethical-substrate approach requires.[13]

Thirdly and finally, let us return briefly to the parable given at the outset. The reader will remember that we imagined a future in which hospital robots are designed to have, or at least simulate, emotions; specifically, empathy. (Recall the posited COLT system.) In general, it seems very hard to deny that human moral reasoning has a strong emotional component, including empathy. For is it not true that one of the reasons humans resist harming their brothers is that they grasp that inflicting such harm causes these others to experience pain? Given this, our logic-based approach to robot moral reasoning (assuming that the human case serves as our touchstone) is admittedly deficient, since no provision has been made for incorporating emotions, or at least computational correlates thereof, into our computational logics. We are currently working on reworking the computational approach to emotions instantiated in (Si et al. 2010) into a logic-based form, after which further augmentation of $\mathcal{DCEC}^*_{CL}$ will be enabled.

see (Arkoudas & Bringsjord 2007). In this approach, traditional program verification is needed only for the one small piece of code that implements proof-checking.

[13] Govindarajulu's (2013) dissertation marks a contribution to the so-called "harder half" of the crowd-sourcing direction. Again, the "easier half," which apparently is what DARPA has hitherto spent money to address, is to use games to allow non-experts playing them to generate *specifications* corresponding to code. The harder half is devoted to proving that such specifications are indeed true with respect to the associated code. In Govindarajulu's novel games, to play is to find proofs that specifications do in fact hold of programs.

References

Arkoudas, K. & Bringsjord, S. (2007), 'Computers, Justification, and Mathematical Knowledge', *Minds and Machines* **17**(2), 185–202.
URL: *http://kryten.mm.rpi.edu/ka_sb_proofs_offprint.pdf*

Bringsjord, S. (2008), '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., Bringsjord, A. & Bello, P. (2013), Belief in the Singularity is Fideistic, *in* A. Eden, J. Moor, J. Søraker & E. Steinhart, eds, 'The Singularity Hypothesis', Springer, New York, NY, pp. 395–408.

Bringsjord, S. & Govindarajulu, N. S. (2013), Toward a Modern Geography of Minds, Machines, and Math, *in* V. C. M¨ller, 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*

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 Coporate Research Environment', *Natural Language Engineering* **10**, 327–348.

Fitting, M. & Mendelsohn, R. L. (1998), *First-order Modal Logic*, Vol. 277, Kluwer, Netherlands.

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

Govindarajulu, N. S. (2013), Uncomputable Games: Games for Crowdsourcing Formal Reasoning, PhD thesis, Rensselaer Polytechnic Institute.

Greco, G., Greco, S. & Zumpano, E. (2003), 'A Logical Framework for Querying and Repairing Inconsistent Databases', *Knowledge and Data Engineering, IEEE Transactions on* **15**(6), 1389–1408.

Hardegree, G. (2011), Introduction to Modal Logic. This is an on-line textbook available, as of February 2012, at this url:
http://people.umass.edu/gmhwww/511/text.htm.

Klein, G. (2010), A Formally Verified OS Kernel. Now What?, *in* M. Kaufmann & L. C. Paulson, eds, 'Interactive Theorem Proving', Vol. 6172 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 1–7.

Klein, G., Elphinstone, K., Heiser, G., Andronick, J., Cock, D., Derrin, P., Elkaduwe, D., Engelhardt, K., Kolanski, R., Norrish, M., Sewell, T., Tuch, H. & Winwood, S. (2009), seL4: Formal Verification of an OS Kernel, *in* 'Proceedings of the ACM SIGOPS 22nd Symposium on Operating systems principles', SOSP '09, ACM, New York, NY, USA, pp. 207–220.

McNamara, P. (2010), Deontic logic, *in* E. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2010 edn. The section of the article discussing a dyadic system is available at:
http://plato.stanford.edu/entries/logic-deontic/chisholm.html.

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.

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.

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'.

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

Stickel, M. E. (2008), 'SNARK - SRI's New Automated Reasoning Kit'. Retrieved on July 26, 2013. http://www.ai.sri.com/~stickel/snark.html.
URL: *http://www.ai.sri.com/ stickel/snark.html*