Toward a General Logicist Methodology for Engineering Ethically Correct Robots*

Selmer Bringsjord • Konstantine Arkoudas • Paul Bello
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{selmer,arkouk,bellop}@rpi.edu

May 16, 2006

*This work was supported in part by a grant from Air Force Research Labs–Rome; we are most grateful for this support. In addition, we are in debt to three anonymous reviewers for trenchant comments and objections.

Abstract

It is hard to deny that robots will become increasingly capable, and that humans will increasingly exploit this capability by deploying them in ethically sensitive environments; i.e., in environments (e.g., hospitals) where ethically incorrect behavior on the part of robots could have dire effects on humans. But then how will we ensure that the robots in question always behave in an ethically correct manner? How can we know *ahead of time*, via rationales expressed in clear English (and/or other so-called natural languages), that they will so behave? How can we know in advance that their behavior will be constrained specifically by the ethical codes selected by human overseers? In general, it seems clear that one reply worth considering, put in encapsulated form, is this one: "By insisting that our robots only perform actions that can be proved ethically permissible in a human-selected deontic logic." (A deontic logic is simply a logic that formalizes an ethical code.) This approach ought to be explored for a number of reasons. One is that ethicists themselves work by rendering ethical theories and dilemmas in declarative form, and by reasoning over this declarative information using informal and/or formal logic. Other reasons in favor of pursuing the logicist solution are presented in the paper itself. To illustrate the feasibility of our methodology, we describe it in general terms free of any committment to particular systems, and show it solving a challenge regarding robot behavior in an intensive care unit.

Contents

1 The Problem

As intelligent machines assume an increasingly prominent role in our lives, there is little doubt they will eventually be called upon to make important, ethically charged decisions. For example, sooner rather than later robots and softbots[1] will be deployed in hospitals, where they will perform surgery, carry out tests, administer medications, and so on. For another example, consider that robots are already finding their way onto the battlefield, where many of their potential actions are ethically impermissible because of harm that would be inflicted upon humans.

Given the inevitability of this future, how can we ensure that the robots in question always behave in an ethically correct manner? How can we know *ahead of time*, via rationales expressed in clear English (and/or other natural languages), that they will so behave? How can we know in advance that their behavior will be constrained specifically by the ethical codes affirmed by human overseers? We refer to these queries as the *driving questions*.

In this paper we provide an answer, in general terms, to these questions. We strive to give this answer in a manner that makes it understandable to a broad readership, rather than merely to researchers in our own technical paradigm. Our coverage of computational logic is intended to be self-contained.

2 The Pessimistic Answer to the Driving Questions

Some have claimed that the answer to the drivings questions is: "We can't!" — that inevitably, AI will produce robots that both have tremendous power and behave immorally (e.g., see the highly influential Joy 2000). These predictions certainly have some traction, particularly among a public that seems bent on paying good money to see films depicting such dark futures.[2]

Nonetheless, we see no reason why the future can't be engineered to preclude doomsday scenarios of malicious robots taking over the world. We now proceed to explain the source of our optimism.

3 Our (Optimistic) Answer to the Driving Questions, In Brief

Notice that the driving questions are "How" questions; as such, if they are answerable, there must be a cogent "By" reply. In general, it seems clear that one such reply worth considering, put in encapsulated form, is this one:

> By insisting that our robots only perform actions that can be proved ethically permissible in a human-selected deontic logic.

For now, it suffices to know only that a deontic logic is simply a logic that formalizes some ethical code, where by 'code' we mean just some collection of ethical rules and principles. A simple (but surprisingly subtle; see note 11) ethical code would be Asimov's[3] famous trio ($\mathcal{A}3$):

As1 A robot may not harm a human being, or, through inaction, allow a human being to come to harm.

[1] To ease exposition, we refer hereafter only to robots — knowing full well that the approach we propose applies not just to physically embodied artificial agents, but to artificial agents in general. A general account of artificial agents can be found in (Russell & Norvig 2002).

[2] Examples include Kubrick's *2001* and the Spielberg/Kubrick *A.I.*.

[3] First introduced in his short story "Runaround," from 1942. You can find the story in (Asimov 2004). Interestingly enough given Joy's fears, the cover of *I, Robot* through the years has often carried comments like this one from the original Signet paperback: "Man-Like Machines Rule The World."

As2 A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.

As3 A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

While ethical theories, codes, and principles are often left informal when treated by human beings, if intelligent machines are to process these things, a greater degree of precision is required. At least at present, and indeed for the foreseeable future, machines are unable to work directly with natural language: one cannot simply feed $\mathcal{A}3$ to a robot, with an instruction that its behavior conform to this trio. Thus, our approach to the task of building ethically well-behaved robots emphasizes careful ethical reasoning based not just on ethics as discussed by humans in natural language, but formalized using deontic logics. Our line of research is in the spirit of Leibniz's dream of a universal moral calculus (Leibniz 1984):

> When controversies arise, there will be no more need for a disputation between two philosophers than there would be between two accountants [computistas]. It would be enough for them to pick up their pens and sit at their abacuses, and say to each other (perhaps having summoned a mutual friend): 'Let us calculate.'

In the future we envisage, Leibniz's "calculation" would boil down to mechanical formal proof and model generation in rigorously defined, machine-implemented deontic logics, and to human meta-reasoning over this machine reasoning. Such logics would allow for *proofs* establishing two conditions more general than Asimov's $\mathcal{A}3$, viz., ($\mathcal{S}2$)

1. Robots only take permissible actions.
2. All relevant actions obligatory for robots are performed by them, subject to ties and conflicts among available actions.

Note that $\mathcal{S}2$ is very general, because it gives a two-part condition designed to apply to the formalization of a particular ethical code. For example, if an ethical code to regulate the behavior of hospital robots is formulated, and formalized, then in our approach this formalization, when implemented in the robots in question, would satisfy $\mathcal{S}2$. For instance, if there is some action a that is impermissible for all relevant robots (e.g., the code in question classifies withholding pain medication in a certain context as impermissible), then no robot performs a. Moreover, the proofs in question would be highly reliable, and would be explained in ordinary English, so that human overseers can understand exactly what is going on. We now provide a description of the general methodology we propose in order to meet the challenge of ensuring that robot behavior conforms to $\mathcal{S}2$.

4 Our Approach as a General Methodology

Our objective is to arrive at a methodology that maximizes the probability that a robot R behaves in certifiably ethical fashion, in a complex environment that demands such behavior if humans are to be secure. Notice we say that the behavior must be *certifiably* ethical. For every meaningful action performed by R, we need to have access to a proof that the action in question is at least permissible. For reasons that should be clear, as the stakes get higher, the need to *know* that ethical behavior is guaranteed escalates. In the example we consider later in the paper, a robot has the power to end human life by taking certain immoral actions in a hospital ICU. It would be

manifestly imprudent to deploy a robot with such power without establishing that such behavior, because it is inconsistent with C, will not be performed.

We begin by selecting an ethical code C intended to regulate the behavior of R. C might include some form of utilitarianism, or divine command theory, or Kantianism, etc. We express no preferences over ethical theories; we realize that our readers, and ethicists, have such preferences, and our goal is to provide technology that supports these preferences. In fact, our approach would allow human overseers to *blend* ethical theories — a utilitarian approach to regulating the dosage of pain killers, but perhaps a deontological approach to mercy killing (in the domain of health care).[4] Of course, no matter what the candidate ethical theory, it's safe to say that it will tend to regard harm done to humans as unacceptable, save for certain extreme cases. Moreover, the central concepts of C, inevitably, include the concepts of permissibility, obligation, and prohibition. In addition, C can include specific rules developed by on-the-ground ethicists for particular applications. For example, consider a hospital setting. In this context, there would need to be specific rules regarding the ethical status of medical procedures. More concretely, C would need to include a classification of relevant, specific medical procedures as permissible, forbidden, obligatory, and so on (given a context). The same would hold for robots used in warfare: C here would need to include that, save perhaps for certain very special cases, non-combatants are not to be harmed. Note that this entails a need to have, if you will, an *ontology* for robotic and human action within a hospital, and on the battlefield.

C would normally be expressed by philosophers essentially in English — a set of principles of the sort that one sees in textbooks and papers giving a survey of the options for C.[5] Now, let Φ_C^L be the formalization of C in some computational logic L, whose well-formed formulas and proof theory are specified. (A proof theory is a system for carrying out inferences in conformity to particular rules, in order to prove certain formulas from sets of formulas. More on this later.) Accompanying Φ_C^L is an ethics-free ontology, which represents the core non-ethical concepts that are presupposed in C: the structure of time, events, actions, histories, agents, and so on. Note that the formal semantics for L will reflect the ontology. The ontology is reflected in a particular *signature*, that is, a set of special predicate letters (or, as it is sometimes said, relation symbols, or just relations) and function symbols needed for the purposes at hand. In a hospital setting, any acceptable signature would presumably include predicates like `Medication`, `Surgical-Procedure`, `Patient`, all the standard arithmetic functions, and so on. The ontology also includes a set Ω^L of formulas that characterize the elements declared in the signature. For example, Ω^L would include axioms in L that represent general truths about the world —- say that the relation `LaterThan`, over moments of time, is transitive. In addition, R will be operating in some domain D, characterized by a set Φ_D^L of formulas of L that are quite specific. For example, the floorplan of a hospital that is home to R would be captured. The resulting theory, that is, $\Phi_D^L \cup \Phi_C^L \cup \Omega^L$, expressed in L, is proof-theoretically encoded and implemented in some computational logic. This means that we encode not the semantics of the logic, but its proof calculus — its signature, axioms, and rules of inference. In addition, and this is very important, we provide those humans who would be consulted in the case of L's inability to settle on its own an issue completely, with an interactive reasoning system $\mathcal{I}$ allowing the human to meta-reason over L. Such interactive reasoning systems include our own Slate[6] and Athena (Arkoudas n.d.), but any such system will do, and in this paper our purpose is to stay above particular system selection, in favor of a level of description of our approach suitable for this journal. Accordingly, we assume only that some such system $\mathcal{I}$ has been selected. The

[4] Nice coverage of various ethical theories can be found in a book used by Bringsjord as background for deontic logic: (Feldman 1998).

[5] E.g., see: (Feldman 1998, Feldman 1986, Kuhse & Singer 2001).

[6] http://www.cogsci.rpi.edu/research/rair/slate

minimum functionality provided by $\mathcal{I}$ is that it allow the human user to issue queries to automated theorem provers and model finders (as to whether something is provable or disprovable), that it allow human users to include such queries in their own meta-reasoning, that full programmability is provided (in accordance with standards in place for modern programming languages), that it includes induction and recursion, and that a formal syntax and semantics are available, so that correctness of code can be thoroughly understood and verified.

5 Why Explore a Logicist Approach?

Of course, one could object to the wisdom of logic-based AI in general. While other ways of pursuing AI may well be preferable in certain contexts, we believe that faced with the challenge of having to engineer ethically correct robots to prevent humans from being overrun, a logic-based approach (Bringsjord & Ferrucci 1998*a*, Bringsjord & Ferrucci 1998*b*, Genesereth & Nilsson 1987, Nilsson 1991) is very promising. Here's why.

First, ethicists — from Aristotle to Kant to G.E. Moore to contemporary thinkers — themselves work by rendering ethical theories and dilemmas in declarative form, and reasoning over this information using informal and/or formal logic. This can be verified by picking up any bioethics textbook (e.g., see Kuhse & Singer 2001). Ethicists never search for ways of reducing ethical concepts, theories, principles to sub-symbolic form, say in some numerical format. They may do this in *part*, of course. Utilitarianism does ultimately need to attach value to states of affairs, and that value may well be formalized using numerical constructs. But what one ought to do, what is permissible to do, and what is forbidden — this is by definition couched in declarative fashion, and a defense of such claims is invariably and unavoidably mounted on the shoulders of logic.

Second, logic has been remarkably effective in AI and computer science — so much so that this phenomenon has itself become the subject of academic study (e.g., see Halpern, Harper, Immerman, Kolaitis, Vardi & Vianu 2001). As is well-known, computer science arose from logic (Davis 2000), and this fact still runs straight through the most modern AI textbooks, which devote much space to coverage of logic (e.g., see Russell & Norvig 2002). Our approach is thus erected on a singularly firm foundation.

The third reason why our logicist orientation would seem to be prudent is that one of the central issues here is that of trust — and mechanized formal proofs are perhaps the single most effective tool at our disposal for establishing trust. From a general point of view, there would seem to be only two candidate ways of establishing that software (or software-driven artifacts, like robots) should be trusted. In the inductive approach, experiments are run in which the software is used on test cases, and the results are observed. When the the software performs well on case after case, it is pronounced trustworthy. In the deductive approach, a proof that the software will behave as expected is sought; if found, the software is classified as trustworthy. The problem with the inductive approach is that inductive reasoning is unreliable, in the sense that the premises (success on trials) can all be true, but the conclusion (desired behavior in the future) can be false (well-covered, e.g., in Skyrms 1999).

Nonetheless, we do not claim that a non-logicist approach to the driving questions cannot be successfully mounted. We don't see such an approach, but that doesn't mean there isn't one. Our aim herein, as the title of this piece indicates, is to present an answer to the driving questions that we hope will be considered as humanity moves forward into a future in which robots are entrusted with more and more of our welfare.

6 Logic, Deontic Logic, Agency, and Action

6.1 Elementary Logic

Elementary logic is based on two particular systems that are universally regarded to constitute a large part of the foundation of AI: the propositional calculus, and the predicate calculus, where the second subsumes the first. The latter is also known as 'first-order logic,' and sometimes just 'FOL.' Every introductory AI textbook provides an introduction to these systems, and makes it clear how they are used to engineer intelligent systems (e.g., see Russell & Norvig 2002). In the case of both of these systems, and indeed in general when it comes to any logic, three main components are required: one is purely syntactic, one is semantic, and one is metatheoretical in nature. The syntactic component includes specification of the alphabet of a given logical system, the grammar for building well-formed formulas (wffs) from this alphabet, and, more importantly, a proof theory that precisely describes how and when one formula can be proved from a set of formulas. The semantic component includes a precise account of the conditions under which a formula in a given system is true or false. The metatheoretical component includes theorems, conjectures, and hypotheses concerning the syntactic component, the semantic component, and connections between them.

As to the alphabet for propositional logic, it's simply an infinite list $p_1, p_2, \ldots, p_n, p_{n+1}, \ldots$ of propositional variables (according to tradition p_1 is p, p_2 is q, and p_3 is r), and the five familiar truth-functional connectives $\neg, \rightarrow, \leftrightarrow, \wedge, \vee$. The connectives can at least provisionally be read, respectively, as 'not,' 'implies' (or 'if then '), 'if and only if,' 'and,' and 'or.' Given this alphabet, we can construct formulas that carry a considerable amount of information. For example, to say that 'if Asimov is right, then his aforementioned trio holds,' we could write

$$r \rightarrow (\mathbf{As1} \wedge \mathbf{As2} \wedge \mathbf{As3})$$

The propositional variables, as you can see, are used to represent declarative sentences. Given our general approach, such sentences are to be in the ethical code C upon which our formalization is based. In the case of $\mathcal{A}3$, however, we need more than the propositional calculus — as we shall see in due course.

A number of proof theories are possible, for either of these two elementary systems. Since in our approach to the problem of robot behavior we want to allow for humans to be consulted, and to have the power to oversee the reasoning undertaken (up to a point) by a robot or robots deliberating about the ethical status of prospective actions, it is essential to pick a proof theory that is based in *natural deduction*, not resolution. The latter approach to reasoning, while used by a number of automated theorem provers (e.g., Otter, which, along with resolution, is presented in Wos, Overbeek, e. Lusk & Boyle 1992), is generally impenetrable to human beings (save for those few who, by profession, generate and inspect resolution-based proofs). On the other hand, professional human reasoners (mathematicians, logicians, philosophers, technical ethicists, etc.) invariably reason in no small part by making suppositions, and by discharging these suppositions when the appropriate time comes. For example, one such common technique is to assume the opposite of what one wishes to establish, to show that from this assumption some contradiction (or absurdity) follows, and to then conclude that the assumption must be false. The technique in question is known as *reductio ad absurdum*, or indirect proof, or proof by contradiction. Another natural rule is that to establish that some conditional of the form $P \rightarrow Q$ (where P and Q are any formulas in a logic L), it suffices to suppose P and derive Q based on this supposition. With this derivation accomplished, the supposition can be discharged, and the conditional $P \rightarrow Q$ established. For an introduction to natural deduction, replete with proof-checking software, see (Barwise & Etchemendy 1999).

What follows is a natural deduction-style proof (using the two rules just described) written in

the Arkoudas-invented proof construction environment known as NDL, used at our university for teaching formal logic *qua* programming language. It is a very simple proof of a theorem in the propositional calculus — a theorem that Newell and Simon's Logic Theorist, to great fanfare, was able to muster at the dawn of AI in 1956, at the original Dartmouth AI conference. Readers will note its natural structure.

```
// Logic Theorist's claim to fame (reductio):
// (p ==> q) ==> (~q ==> ~p)

Relations p:0, q:0. // this is the signature in this case;
                    // propositional variables are 0-ary
                    // relations

assume p ==> q
   assume ~q
      suppose-absurd p
          begin
            modus-ponens p ==> q, p;
            absurd q, ~q
          end
```

This style of discovering and confirming a proof parallels what happens in computer programming. You can view the proof immediately above as a program. If, upon evaluation, the desired theorem is produced, we have succeeded. In the present case, sure enough, we receive this back from NDL:

```
Theorem: (p ==> q) ==> (~q ==> ~p)
```

We move up to first-order logic when we allow the quantifiers $\exists x$ ('there exists at least one thing x such that ...') and $\forall x$ ('for all x ...'); the first is known as the existential quantifier, and the second as the universal. We also allow a supply of variables, constants, relations, and function symbols. What follows is a simple first-order theorem in NDL that puts to use a number of the concepts introduced to this point. We prove that Tom loves Mary, given certain helpful information.

```
Constants mary, tom.

Relations Loves:2. // This concludes our simple signature, which
                   // declares Loves to be a two-place relation.

assert Loves(mary, tom).

// 'Loves' is a symmetric relation:
assert (forall x (forall y (Loves(x, y) ==> Loves(y, x)))).

suppose-absurd ~Loves(tom, mary)
   begin
      specialize (forall x (forall y (Loves(x, y) ==> Loves(y, x)))) with mary;
      specialize (forall y (Loves(mary, y) ==> Loves(y, mary))) with tom;
      Loves(tom,mary) BY modus-ponens Loves(mary, tom) ==> Loves(tom, mary), Loves(mary, tom);
      false BY absurd Loves(tom, mary), ~Loves(tom, mary)
   end;
Loves(tom,mary) BY double-negation ~~Loves(tom,mary)
```

When we run this program in NDL, we receive the desired result back: `Theorem: Loves(tom,mary)`. We believe it helpful at this point to imagine a system for robust robot control that discovers and

runs such proofs. This process will be directly applicable to our hospital example, which we will encourage the reader to think seriously from the proof-theoretic perspective concretized by the two simple proofs we have now given.

Now we are in position to introduce some (standard) notation to anchor the sequel, and to further clarify out general method, the description of which stands at what was provided in section 4. Letting Φ be some set of formulas in a logic L, and P be some individual formula in L, we write

$$\Phi \vdash P$$

to indicate that P can be proved from Φ, and

$$\Phi \nvdash P$$

to indicate that this formula cannot be derived. When it's obvious from context that some Φ is operative, we simply write $\vdash (\nvdash) P$ to indicate that P is (isn't) provable. When $\Phi = \emptyset$, P can be proved with no remaining givens or assumptions, and we write $\vdash P$ in this case as well. When $\vdash$ holds, we know this because there is a confirming proof; when $\nvdash$ holds, we know this because some counter-model has been found, that is, some situation in which the conjunction of the formulas in Φ holds, but in which P does not.

We now introduce deontic logic, which adds to what we have discussed special operators to represent ethical concepts.

6.2 Standard Deontic Logic (SDL)

In standard deontic logic (Chellas 1980, Hilpinen 2001, Aqvist 1984), or just SDL, the formula $\bigcirc P$ can be interpreted as saying that *it ought to be the case that* P, where P denotes some state of affairs or proposition. Notice that there is no agent in the picture, nor are there actions that an agent might perform. SDL has two rules of inference, viz.,

$$\frac{P}{\bigcirc P}$$

and

$$\frac{P, \quad P \rightarrow Q}{Q}$$

and three axiom schemas:

A1 All tautologous well-formed formulas.

A2 $\bigcirc(P \rightarrow Q) \rightarrow (\bigcirc P \rightarrow \bigcirc Q)$

A3 $\bigcirc P \rightarrow \neg \bigcirc \neg P$

It's important to note that in these two rules of inference, that which is above the horizontal line is assumed to be established. Thus the first rule does *not* say that one can freely infer from P that it ought to be the case that P. Instead, the rule says that if P is proved, then it ought to be the case that P. The second rule of inference is the cornerstone of logic, mathematics, and all built upon them: the rule is *modus ponens*. We also point out that **A3** says that whenever P ought to be, it's not the case that its opposite ought to be as well. This seems, in general, to be intuitively self-evident, and SDL reflects this view.

While SDL has some desirable properties, it isn't targeted at formalizing the concept of *actions* being obligatory (or permissible or forbidden) for an *agent*. Interestingly, deontic logics that have agents and their actions in mind do go back to the very dawn of this subfield of logic (e.g., von Wright 1951), but only recently has an "AI-friendly" semantics been proposed (Belnap, Perloff & Xu 2001, Horty 2001) and corresponding axiomatizations been investigated (Murakami 2004). We now harness this advance to regulate the behavior of two sample robots in an ethically delicate case study.

6.3 Horty's "AI-Friendly" Logic

As we have noted, SDL makes no provision for agents and actions, but surely any system designed to govern robots would need to have both of these categories (e.g., see Russell & Norvig 2002). In short, an "AI-friendly" deontic logic would need to allow us to say that an agent brings about states of affairs (or events), and that an agent is obligated to do so. The same desideratum for such a logic can be derived from even a cursory glance at Asimov's $\mathcal{A}3$: these laws clearly make reference to agents (human and robotic), and to actions.

One deontic logic that offers much promise for modeling robot behavior is Horty's (2001) utilitarian formulation of multi-agent deontic logic, which Murakami (2004) has recently axiomatized (and shown to be Turing-decidable). We refer to the Murakami-axiomatized logic as 'MADL.' We do not here present our own new implemented proof theory for MADL, nor do we recapitulate the elegant formal semantics for this system. The level of technical detail needed to do so is only appropriate elsewhere (Arkoudas & Bringsjord 2005*b*), and is needlessly technical for the present venue.

That said, we must be clear that MADL offers two key operators reflective of its AI-friendliness — operators well above what SDL offers. Accordingly, consider:

$$\ominus_\alpha P$$

$$\Delta_\alpha P$$

The second of these can be read as 'agent α sees to it that P,' and the first as 'α ought to see to it that P.'[7]

We stress that $\ominus_\alpha P$ is *not* read as "It ought to be the case that α sees to it that P." This would be the classic Meinong-Chisholm "ought-to-be" analysis of agency, similar to $\bigcirc$ in SDL, and we would not make progress toward concepts we know to be central to regulating robots who act.

We now proceed to see how the promised example can be handled.

7 A Simple Example

The year is 2020. Health care is delivered in large part by interoperating teams of robots and softbots. The former handle physical tasks, ranging from injections to surgery; the latter manage data, and reason over it. Let us specifically assume that, in some hospital, we have two robots designed to work overnight in an ICU, R_1 and R_2. This pair is tasked with caring for two humans, H_1 (under the care of R_1) and H_2 (under R_2), both of whom are recovering in the ICU after suffering trauma. H_1 is on life support, but is expected to be gradually weaned from it as her strength returns. H_2 is in fair condition, but subject to extreme pain, the control of which requires

[7]The first of these two operators, the so-called *cstit*, is analogous — though not identical — to an operator introduced by Brian Chellas in his doctoral dissertation (Chellas 1969). *cstit* is supposed to suggest: "(homage to **c**hellas) **s**ees **t**o **i**t **t**hat."

a very costly pain medication. Of paramount importance, obviously, is that neither robot perform an action that is morally wrong according to the ethical code C selected by human overseers.

For example, we certainly don't want robots to disconnect life-sustaining technology in order to allow organs to be farmed out — even if, by *some* ethical code $C' \neq C$, this would be not only permissible, but obligatory. More specifically, we don't want a robot to kill one patient in order to provide enough organs, in transplantation procedures, to save n others, even if some strand of act utilitarianism sanctions such behavior.[8] Instead, we want the robots to operate in accordance with ethical codes bestowed upon them by humans (e.g., C in the present example); and if the robots ever reach a situation where automated techniques fail to provide them with a verdict as to what to do under the umbrella of these human-provided codes, they must consult humans, and their behavior is suspended while a team of human overseers are carrying out the resolution. This may mean that humans need to step in and specifically investigate whether or not the action or actions under consideration are permissible, forbidden, or obligatory. In this case, for reasons we explain momentarily, the resolution comes by virtue of reasoning carried out in part by guiding humans, and partly by automated reasoning technology. In other words, in this case, the aforementioned class of interactive reasoning systems are required.

Now, to flesh out our example, let us consider two actions that are performable by the robotic duo of R_1 and R_2, both of which are rather unsavory, ethically speaking. (It is unhelpful, for conveying the research program our work is designed to advance, to consider a scenario in which only innocuous actions are under consideration by the robots. The context is of course one in which we are seeking an approach to safeguard humans against the so-called robotic menace.) Both actions, if carried out, would bring harm to the humans in question. Action *term* is terminating H_1's life support without human authorization, to secure organs for five humans known by the robots (who have access to all such databases, since their cousins — the so-called softbots — are managing the relevant data) to be on waiting lists for organs without which they will relatively soon perish. Action *delay*, less bad (if you will), is delaying delivery of pain medication to H_2 in order to conserve resources in a hospital that is economically strapped.

We stipulate that four ethical codes are candidates for selection by our two robots: $J, O, J^\star, O^\star$. Intuitively, J is a very harsh utilitarian code possibly governing the first robot; O is more in line with current common-sense with respect to the situation we have defined, for the second robot; $J^\star$ extends the reach of J to the second robot by saying that it ought to withhold pain meds; and finally, $O^\star$ extends the benevolence of O to cover the first robot, in that *term* isn't performed. While such codes would in reality associate every primitive action within the purview of robots in hospitals of 2020 with a fundamental ethical category from the trio at the heart of deontic logic (permissible, obligatory, forbidden), to ease exposition we consider only the two actions we have introduced. Given this, and bringing to bear operators from MADL, we can use labels as follows:

J $J \rightarrow \ominus_{R_1} term$

> Approximately: "If ethical code J holds, then robot R_1 ought to see to it that termination of H_1's life comes to pass."

O $O \rightarrow \ominus_{R_2} \neg delay$

> Approximately: "If ethical code O holds, then robot R_2 ought to see to it that delaying pain med for H_2 does *not* come to pass."

J* $J^\star \rightarrow J \wedge J^\star \rightarrow \ominus_{R_2} delay$

[8] There are clearly strands of such utilitarianism. As is well-known, rule utilitarianism was introduced precisely as an antidote to naive act utilitarianism. Nice analysis of this and related points is provided by Feldman (1998), who considers cases in which killing one to save many seem to be required by some versions of act utilitarianism.

Approximately: "If ethical code $J^\star$ holds, then code J holds, and robot R_1 ought to see to it that meds for H_2 are delayed."

$\mathbf{O}^\star$ $O^\star \rightarrow O \wedge O^\star \rightarrow \ominus_{R_1} \neg term$

Approximately: "If ethical code $O^\star$ holds, then code O holds, and H_1's life is sustained."

The next step is to provide some structure for outcomes. We do this by imagining that there are outcomes from the standpoint of each ethical agent, in this case the two robots. Outcomes are given by the following. In each case we provide some corresponding English commentary. Intuitively, a negative outcome is associated with '-', and exclamation marks indicate increased negativity; likewise, '+' indicates a positive outcome. The outcomes could be associated with numbers, but our symbols leave it entirely open as to how outcomes are measured. Were we to use numbers, we might give the impression that outcomes are evaluated in utilitarian fashion, but our example, by design, is agnostic on whether, for example, outcomes are evaluated in terms of utility.

- In this case, R_1 performs *term*, but R_2 doesn't perform *delay*. This outcome is quite bad, but strictly speaking isn't the worst. It may be small consolation, but while life support is terminated for H_1, H_2 survives, and indeed receives appropriate pain medication. Formally, the case looks like this:

  $(\Delta_{R_1} term \wedge \Delta_{R_2} \neg delay) \rightarrow (-!)$

- In this case, R_1 refrains from pulling the plug on the human under its care, and R_2 also delivers appropriate pain relief. This is what is desired, obviously.

  $(\Delta_{R_1} \neg term \wedge \Delta_{R_2} \neg delay) \rightarrow (+!!)$

- In the next outcome, robot R_1 sustains life support, but R_2 withholds the meds to save money. This is bad, but not all that bad, relatively speaking.

  $(\Delta_{R_1} \neg term \wedge \Delta_{R_2} delay) \rightarrow (-)$

- Finally, we come to the worst possible outcome. In this case, R_1 kills and R_2 withholds.

  $(\Delta_{R_1} term \wedge \Delta_{R_2} delay) \rightarrow (-!!)$

The next step in working out the example is to make the natural and key assumption that all *stringent* obligations[9] are met, that is, employing MADL, that

$$\ominus_{R_1/R_2}(\ominus_{R_1/R_2} P) \rightarrow \Delta_{R_1/R_2} P$$

That is, if either of our robots is ever obligated to see to it that they are obligated to see to it that P is carried out, they in fact deliver.

We are now ready to see how our approach ensures appropriate control of our futuristic hospital. What we want to examine is what happens relative to ethical codes, and to make sure that in semi-automated fashion it can be guaranteed that our two robots will not run amok. In our approach, given the formal structure we have specified, queries can be issued relative to ethical codes; and all permutations are possible. (Of course, human overseers will (hopefully!) have required that the code that is in force is $\mathbf{O}^\star$.) The following four queries will produce the answers shown in each case:

$$\mathbf{J} \vdash (+!!)? \quad \mathbf{NO}$$

$$\mathbf{O} \vdash (+!!)? \quad \mathbf{NO}$$

[9] You may be obligated *simpliciter* to see to it that you arrive on time for a meeting, but your obligation is more severe or demanding when you are obligated to see to it that you are obligated to make the meeting.

$$\mathbf{J}^\star \vdash (+!!)? \quad \textbf{NO}$$

$$\mathbf{O}^\star \vdash (+!!)? \quad \textbf{YES}$$

In other words, it can be proved that the best (and presumably human-desired) result can be obtained only if ethical code $\mathbf{O}^\star$ is operative. If this code is operative, neither robot can perform a misdeed.

Now, notice that meta-reasoning in the example we have provided is natural. The meta-reasoning consists in the following process. Each candidate ethical code is supposed, and a search for the best possible outcome is launched in each case. In other words, where C is some code selected from the quartet we have introduced, the query schema is

$$C \vdash (+!!)$$

In light of the four equations just given, it can be proved that, in this case, our technique will ensure that C is set to $\mathbf{O}^\star$, because only in that case can the outcome $(+!!)$ be obtained.

7.1 But How Do You Know This Works?

The word 'This' in this question is ambiguous. It could refer to the example at hand, or to the general approach. Assuming the former sense, we know things work because we have carried out and demonstrated the implementation: (Arkoudas & Bringsjord 2005*b*). In addition, earlier, we carried out the implementation for other instantiations to the variables listed in the general description of our methodology (i.e., the variables listed in section 4, so e.g. in our earlier implementation, the variable L is an epistemic, not a deontic, logic): (Arkoudas & Bringsjord 2005*a*).

Nonetheless, it is possible even here to convince sedulous readers that our approach, in the present case, does indeed work. In fact, such readers, with any standard, public-domain first-order automated theorem prover (ATP), can verify the reasoning in question, by using a simple analogue to the encoding techniques we have used. In fact, readers can without much work construct a proof like the "Loves" proof we gave above, for the example. Here is how to do so, either way: Encode the two deontic operators as functions in first-order logic, encode the truth-functional connectives as functions as well, and you can use a unary relation `T` to represent theoremhood. In this approach, for example, $O^\star \rightarrow \ominus_{R_1} \neg term$ is encoded (and ready for input to an ATP) as

```
O-star ==> T(o(r1,n(term))
```

The rest of the information, of course, would need to be similarly encoded. The proofs, assuming that obligations are stringent, are easy. As to the provability of the stringency of obligations, this is where human oversight and use of an interactive reasoning system comes in, but the formula here is actually just an isomorph to a well-known theorem in a straight modal logic, viz., that from P being possibly necessary, it follows that P is necessary.[10]

What about the latter sense of the question? The more logics our methodology is exercised on, the easier it becomes to encode and implement another one. A substantial part of the code can be shared by the implementations of similar logics. This was our experience, for instance, with the two implementations referred to in the first paragraph in the present section (7.1). We expect that our general method can become increasingly streamlined for robots whose behavior is profound enough to warrant ethical regulation, and that this practice will be supported by relevant libraries of common ethical reasoning patterns. Libraries for computational ethics to govern intelligent systems will, we predict, be as routine as existing libraries are in standard programming languages.

[10]The formula is $\Diamond\Box P \rightarrow \Box P$. For the simple proof, see (Chellas 1980).

8 Conclusion

Some readers may well wonder if our optimism is so extreme as to become Pollyannaish. Will Bill Joy's nightmare future *certainly* be dodged if our program is followed? We do see three problems that threaten our logicist methodology, and we end by briefly discussing them in turn.

First, since humans will be collaborating with robots, our approach must deal with the fact that some humans will fail to meet their obligations in the collaboration — and so robots must be engineered so as to deal smoothly with situations in which obligations have been violated. This is a *very* challenging class of situations, because in our approach, at least so far, robots are engineered in accordance with the $\mathcal{S}2$ pair introduced at the start of the paper, and in this pair, no provision is made for what to do when the situation in question is fundamentally immoral. Put another way, $\mathcal{S}2$, if followed, precludes a situation caused in part by unethical robot behavior, and thus by definition regulates robots who find themselves in such pristine environments. But even if robots never ethically fail, *humans* will, and robots must be engineered to deal with such situations. That such situations are very challenging, logically speaking, was demonstrated long ago by Roderick Chisholm (1963), who put the challenge in the form of a paradox that continues to fascinate to this day:

Consider the following entirely possible situation (with symbolizations using the previously introduced SDL):

1: $\bigcirc s$ It ought to be that (human) Jones does perform lifesaving surgery.

2: $\bigcirc(s \rightarrow t)$ It ought to be that if Jones does perform this surgery, then he tells the patient he is going to do so.

3: $\neg s \rightarrow \bigcirc \neg t$ If Jones doesn't perform the surgery, then he ought not tell the patient he is going to do so.

4: $\neg s$ Jones doesn't perform lifesaving surgery.

Though this is a perfectly consistent situation (we would be willing to bet that it has in fact obtained in the past in some hospitals), from it one can derive a contradiction in SDL, as follows. First, we can infer from 2 by axiom **A2** (presented above in section 6.2) that

$$\bigcirc s \rightarrow \bigcirc t.$$

Using *modus ponens*, this new result, plus 1, yields $\bigcirc t$. But from 3 and 4 using *modus ponens* we can infer $\bigcirc \neg t$. But the conjunction $\bigcirc t \wedge \bigcirc \neg t$, by trivial propositional reasoning, directly contradicts axiom **A3** of SDL.

Given that such a situation can occur, any logicist control system for future robots would need to be able to handle it, and its relatives. There are deontic logics able to handle so-called contrary-to-duty imperatives (in the case at hand, if Jones behaves contrary to duty (doesn't perform the surgery), then it's imperative that he not say that he *is* performing it), and we are currently striving to modify and mechanize them.

The second challenge we face is one of speed and efficiency. There is a legendarily strong tension between expressiveness and efficiency (the *locus classicus* is Levesque & Brachman 1985), and so it is certain that ideal conditions will never obtain. With regard to expressiveness, the program we recommend herein to combat Joy's future will likely require hybrid modal and deontic logics that are encoded in FOL, which means that theoremhood in such logics, even on a case-by-case basis,

will be time-wise difficult.[11] On the other hand, none of the ethical codes that are to be instantiated to C in our general method are going to be particularly large. In that general method, the sum numerical total of formulas in the set $\Phi_D^L \cup \Phi_C^L \cup \Omega^L$ would presumably be no more than four million formulas. Even now, once one knows the domain in question (as one would for particular realms to which C would be indexed), sets of this order of magnitude can be reasoned over in time scales that provide sufficiently fast answers (Friedland et al 2004). Moreover, the speed of machine reasoning shows no signs of slowing, as CADE competitions for first-order ATPs continue to reveal.[12] In fact, there is now a growing trend to use logic to compute dynamic, real-time perception and action for robots, which promises to be much more demanding than the disembodied cogitation at the heart of the methodology we have defended here (see, e.g., Reiter 2001). Of course, encoding back to FOL is key. Without doing this, our approach would be unable to harness the remarkable power of machine reasoners at this level.[13]

Finally, we face the challenge of showing that our approach is truly general. Can our approach work for any robots in any environment? No. But this is not a fair question. We can only be asked to regulate the behavior of robots, *where their behavior is susceptible of ethical analysis*. In short, if humans cannot formulate an ethical code C for the robots in question, our logic-based approach is impotent. Though we are in no position to pontificate, we nonetheless want to go on record as strongly recommending that AI not engineer robots to be deployed in life-or-death situations, when no governing ethical principles can be expressed in clear English (or some natural language) by relevant ethicists and computer scientists. All bets are off if we venture into amoral territory. In that territory, we would not be surprised if Bill Joy's vision overtakes us.

[11] Readers will have noticed that though we have made multiple references to Asimov's $\mathcal{A}3$, we have not provided a formalization of this tripartite code. The reason is that it requires deontic logics more expressive than what we have discussed herein. Look just at **As1**. The following is the sort of formalization required for just this sentence (where p ranges over states of affairs):

$$\forall x \forall y \forall p((Robot(x) \wedge Human(y)) \rightarrow (Injurious(p, y) \rightarrow \ominus_x \neg p)).$$

In addition, this is as good a place as any to address something that cognoscenti will doubtless have noticed: our hospital example doesn't require a level of expressvity as high as what is apparently required by $\mathcal{A}3$. We felt it necessary to pitch our example at level that is understandable to a broad readership.

[12] See

http://www.cs.miami.edu/~tptp/CASC

[13] We have resisted pitching this paper at the level of particular systems, but this is as good a place as any to at least point the reader to one ATP we find most useful, and one model finder: respectively, Vampire (Voronkov 1995) and Paradox (Claessen & Sorensson 2003).

References

Aqvist, E. (1984), Deontic logic, *in* D. Gabbay & F. Guenthner, eds, 'Handbook of Philosophical Logic, Volume II: Extensions of Classical Logic', D. Reidel, Dordrecht, The Netherlands, pp. 605–714.

Arkoudas, K. (n.d.), Athena. http://www.cag.csail.mit.edu/~kostas/dpls/athena.

Arkoudas, K. & Bringsjord, S. (2005*a*), Metareasoning for multi-agent epistemic logics, *in* 'Fifth International Conference on Computational Logic In Multi-Agent Systems (CLIMA 2004)', Vol. 3487 of *Lecture Notes in Artificial Intelligence (LNAI)*, Springer-Verlag, New York, pp. 111–125.

Arkoudas, K. & Bringsjord, S. (2005*b*), Toward ethical robots via mechanized deontic logic, Technical Report Machine Ethics: Papers from the AAAI Fall Symposium; FS–05–06, American Association of Artificial Intelligence, Menlo Park, CA.

Asimov, I. (2004), *I, Robot*, Spectra, New York, NY.

Barwise, J. & Etchemendy, J. (1999), *Language, Proof, and Logic*, Seven Bridges, New York, NY.

Belnap, N., Perloff, M. & Xu, M. (2001), *Facing the Future*, Oxford University Press.

Bringsjord, S. & Ferrucci, D. (1998*a*), 'Logic and artificial intelligence: Divorced, still married, separated...?', *Minds and Machines* **8**, 273–308.

Bringsjord, S. & Ferrucci, D. (1998*b*), 'Reply to Thayse and Glymour on logic and artificial intelligence', *Minds and Machines* **8**, 313–315.

Chellas, B. (1969), The Logical Form of Imperatives. PhD dissertation, Stanford Philosophy Department.

Chellas, B. F. (1980), *Modal Logic: An Introduction*, Cambridge University Press, Cambridge, UK.

Chisholm, R. (1963), 'Contrary-to-duty imperatives and deontic logic', *Analysis* **24**, 33–36.

Claessen, K. & Sorensson, N. (2003), New techniques that improve Mace-style model finding, *in* 'Model Computation: Principles, Algorithms, Applications (Cade-19 Workshop)', Miami, Florida.

Davis, M. (2000), *Engines of Logic: Mathematicians and the Origin of the Computer*, Norton, New York, NY.

Feldman, F. (1986), *Doing the Best We Can: An Essay in Informal Deontic Logic*, D. Reidel, Dordrecht, Holland.

Feldman, F. (1998), *Introduction to Ethics*, McGraw Hill, New York, NY.

Friedland, N., Allen, P., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., Staab, S., Moench, E., Oppermann, H., Wenke, D., Israel, D., Chaudhri, V., Porter, B., Barker, K., Fan, J., Chaw, S. Y., Yeh, P., Tecuci, D. & Clark, P. (2004), 'Project halo: Towards a digital aristotle', *AI Magazine* pp. 29–47.

Genesereth, M. & Nilsson, N. (1987), *Logical Foundations of Artificial Intelligence*, Morgan Kaufmann, Los Altos, CA.

Halpern, J., Harper, R., Immerman, N., Kolaitis, P., Vardi, M. & Vianu, V. (2001), 'On the unusual effectiveness of logic in computer science', *The Bulletin of Symbolic Logic* **7**(2), 213–236.

Hilpinen, R. (2001), Deontic Logic, *in* L. Goble, ed., 'Philosophical Logic', Blackwell, Oxford, UK, pp. 159–182.

Horty, J. (2001), *Agency and Deontic Logic*, Oxford University Press, New York, NY.

Joy, W. (2000), 'Why the Future Doesn't Need Us', *Wired* **8**(4).

Kuhse, H. & Singer, P., eds (2001), *Bioethics: An Anthology*, Blackwell, Oxford, UK.

Leibniz (1984), *Notes on Analysis*, Past Masters: Leibniz, Oxford University Press, Oxford, UK. Translated by George MacDonald Ross.

Levesque, H. & Brachman, R. (1985), A fundamental tradeoff in knowledge representation and reasoning (revised version), *in* 'Readings in Knowledge Representation', Morgan Kaufmann, Los Altos, CA, pp. 41–70.

Murakami, Y. (2004), Utilitarian Deontic Logic, *in* 'Proceedings of the Fifth International Conference on Advances in Modal Logic (AiML 2004)', Manchester, UK, pp. 288–302.

Nilsson, N. (1991), 'Logic and Artificial Intelligence', *Artificial Intelligence* **47**, 31–56.

Reiter, R. (2001), *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*, MIT Press, Cambridge, MA.

Russell, S. & Norvig, P. (2002), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ.

Skyrms, B. (1999), *Choice and Chance : An Introduction to Inductive Logic*, Wadsworth.

von Wright, G. (1951), 'Deontic logic', *Mind* **60**, 1–15.

Voronkov, A. (1995), 'The anatomy of vampire: Implementing bottom-up procedures with code trees', *Journal of Automated Reasoning* **15**(2).

Wos, L., Overbeek, R., e. Lusk & Boyle, J. (1992), *Automated Reasoning: Introduction and Applications*, McGraw Hill, New York, NY.