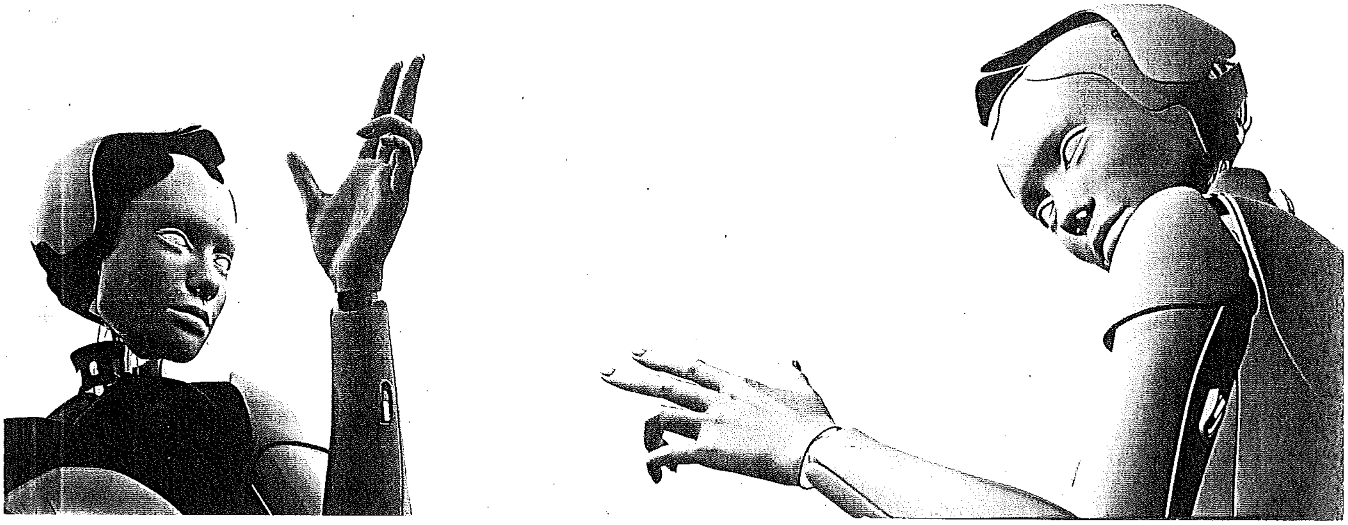
# ROBOT ETHICS

## THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS

EDITED BY

Patrick Lin, Keith Abney, and
George A. Bekey



# Robot Ethics

## The Ethical and Social Implications of Robotics

Edited by Patrick Lin, Keith Abney, and George A. Bekey

The MIT Press
Cambridge, Massachusetts
London, England

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email special_sales@mitpress.mit.edu or write to Special Sales Department, The MIT Press, 55 Hayward Street, Cambridge, MA 02142.

This book was set in Stone by Toppan Best-set Premedia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Robot ethics : the ethical and social implications of robotics / edited by Patrick Lin,
Keith Abney, and George A. Bekey.
p. cm.—(Intelligent robotics and autonomous agents series)
Includes bibliographical references and index.
ISBN 978-0-262-01666-7 (hardcover : alk. paper) 1. Robotics—Human factors. 2. Robotics—Moral and ethical aspects. 3. Robotics—Social aspects. 4. Robots—Design and construction.
I. Lin, Patrick. II. Abney, Keith, 1963– III. Bekey, George A., 1928–
TJ211.49.R62 2012
174′.9629892—dc23
2011016639

10 9 8 7 6 5 4 3 2 1

# 6 The Divine-Command Approach to Robot Ethics

Selmer Bringsjord and Joshua Taylor

Perhaps it is generally agreed that robots on the battlefield, especially those with lethal power, should be ethically regulated. But, then, in what should such regulation consist? Presumably, in the fact that all the significant actions performed by such robots are in accordance with some ethical code. But, of course, the question arises as to *which* code. One narrow option is that the code is a set of *rules of engagement* affirmed by some nation or group; this approach, described later in this chapter, has been taken by Arkin (2008, 2009).[1] Another is utilitarian, represented in computational deontic logic, as explained, for instance, by Bringsjord, Arkoudas, and Bello (2006), and summarized here. Yet another is likewise based on computational logic, but using a logic that captures some other mainstream ethical theory (e.g., Kantian deontology, or Ross's "right mix" direction); this possibility has been rigorously pursued by Anderson and Anderson (2006; Anderson, Anderson, and Armen 2008). But there is a radically different possibility that hitherto hasn't arrived on the scene: the controlling moral code could be viewed as coming straight from God. There is some very rigorous work along this line, known as "divine-command ethics." In a world where human fighters and the general populations supporting them often see themselves as championing God's will in war, divine-command ethics is quite relevant to military robots. Put starkly, on a planet where so-called holy wars are waged time and time again under a generally monotheistic scheme, it seems more than peculiar that heretofore robot ethics (or "roboethics") has been bereft of the systematic study of such ethics on the basis of monotheistic conceptions of what is morally right and wrong. This chapter introduces divine-command ethics in the form of the computational logic *LRT**, intended to eventually be suitable for regulating a real-world warfighting robot. Our work falls in general under the approach to engineering AI systems on the basis of formal logic (Bringsjord 2008c).

The chapter is structured as follows. We first set out the general context of roboethics in a military setting (section 6.1), and point out that the divine-command approach has been absent. We then introduce the divine-command computational logic *LRT** (section 6.2), concluding this section with a scenario in which a robot is constrained by dynamic use of the logic. We end (section 6.3) with some remarks about next steps in the divine-command roboethics program.

## 6.1 The Context for Divine-Command Roboethics

There are several branches of ethics. A standard tripartite breakdown splits the field into *metaethics*, *applied ethics*, and *normative ethics*. The second and third branches directly connect to our roboethics R&D; we discuss the connection immediately after briefly summarizing the trio. For more detailed coverage, the reader is directed to Feldman (1978), which conforms with arguably the most sophisticated published presentation of utilitarianism from the standpoint of the semantics of deontic logic (Feldman 1986). Much of our prior R&D has been based on this same deontic logic (e.g., Bringsjord, Arkoudas, and Bello 2006).

*Metaethics* tries to determine the ontological status of the basic concepts in ethics, such as *right* and *wrong*. For example, are matters of morals and ethics more like matters of fact or of opinion? Who determines whether something is good or bad? Is there a divine being who stipulates what is right or wrong, or a Platonic realm that provides truth-values to ethical claims, independently of what anyone thinks? Is ethics merely *in the head,* and if so, how can any one moral outlook be seen as *better* than any other? As engineers bestowing ethical qualities to robots (in a manner soon to be explained), we are automatically confronted with these metaethical issues, especially given the power to determine a robot's *sense* of right and wrong. Is this an arbitrary choice of the programmer, or are there objective guidelines to determine whether the moral outlook of one robot is better than that of any other robot or, for that matter, of a human? Reflecting on these issues with regard to robots, one quickly gains an appreciation of these important questions, as well as a perspective to potentially answer them. Such reflection is an inevitable consequence of the engineering that is part and parcel of practical roboethics.

*Applied ethics* is more practical and specific. Applied ethics *starts* with a certain set of moral guides, and then applies them to specific domains so as to address specific moral dilemmas arising therein. Thus, we have such disciplines as bioethics, business ethics, environmental ethics, engineering ethics, and many others. A book written by one of us in the past can be viewed as following squarely under bioethics (Bringsjord 1997). Given that robots have the potential to interact with us and our environment in complex ways, the practice of building robots quickly raises all kinds of applied ethical questions: what potential harmful consequences may come from the building of these robots? What happens to important moral notions such as autonomy and privacy when robots are starting to become an integral part of our lives? While many of these issues overlap with other fields of engineering, the potential of robots to become ethical agents themselves raises an additional set of moral questions, including: do such robots have any rights and responsibilities?

"Normative ethics," or "moral theory," compares and contrasts ways to define the concepts "obligatory," "forbidden," "permissible," and "supererogatory." Normative ethics investigates which actions we ought to, or ought not to, perform, and why. "Consequentialist" views render judgments on actions depending on their outcomes, while "nonconsequentialist" views consider the intent behind actions, and thus the inherent duties, rights, and responsibilities that may be involved, independent of particular outcomes. Well-known consequentialist views include egoism, altruism, and utilitarianism; the best-known nonconsequentialist view is probably Kant's theory of moral behavior, the kernel of which is that people should never be treated as a means to an end.

### 6.1.1 Where Our Work Falls

Our work mainly falls within normative ethics, and in two important ways. First, given any particular normative theory *T*, we take on the burden of finding a way to engineer a robot with that particular outlook by deriving and specializing from *T* a particular ethical code *C* that fits the robot's environment, and of *guaranteeing* that a lethal robot does indeed adhere to it. Second, robots infused with ethical codes can be placed under different conditions to see how different codes play out. Strengths and weaknesses of the ethical codes can be observed and empirically studied; this may inform the field of normative ethics. Our work also lies between metaethics and applied ethics. Like metaethics, our primary concern is not with specific moral dilemmas, but rather with general theories and their application to any domain. Like applied ethics, we do not ask for the deep metaphysical status of any of these theories, but rather take them as they are, and consider their outcomes in applications.

### 6.1.2 The Importance of Robot Ethics

Joy (2000) has famously predicted that the future will bring our demise, in no small part because of advances in AI and robotics. While Bringsjord (2008b) rejects this fatalism, if we assume that robots in the future will have more and more autonomy and lethal power, it seems reasonable to be concerned about the possibility that what is now fiction from Asimov, Kubrick, Spielberg, and others, will become morbid reality. However, the importance of engineering ethically correct robots does not derive simply from what creative writers and futurists have written. The U.S. defense community now openly and aggressively affirms the importance of such engineering. A recent extensive and enlightening survey of the overall landscape is provided by Lin, Bekey, and Abney (2008), in their thorough report prepared for the Office of Naval Research, U.S. Department of the Navy, in which the possibility and need of creating ethical robots is analyzed. Their recommended goal is not to make fully ethical machines, but simply machines that perform better than humans in isolated cases. Lin, Bekey, and Abney conclude that the risks and potential negatives of perfectly ethical robots are greatly overshadowed by the benefits they would provide over human peacekeepers and warfighters and thus should be pursued.

We are more pessimistic. While human warfighters remotely control the robots discussed in Lin, Bekey, and Abney (2008), the Department of Defense's Unmanned Systems Integrated Roadmap supports the desire for increasing autonomy. We view the problem as follows: gradually, because of economic and social pressures that will be impossible to suppress, and are already in play, autonomous warfighting robots with lethal power will be deployed in all theaters of war. For example, where defense and social programs expenditures increasingly outstrip revenues from taxation, cost cutting via removing expensive humans from the loop will prove irresistible. Humans are still firmly in the "kill chain" today, but their gradual removal in favor of inexpensive and expendable robots is inevitable. Even if our pessimism were incorrect, only those with Pollyanna-like views of the future would resist our call to at least plan for the *possibility* that this dark outcome may unfold; such prudent planning sufficiently motivates the roboethical engineering we call for.

### 6.1.3 Necessary and Sufficient Conditions for an Ethically Correct Robot

The engineering antidote is to ensure that tomorrow's robots reason in correct fashion with the ethical codes selected. A bit more precisely, we have *ethically correct* robots when they satisfy the following three *core desiderata*.[2]

**D1** Robots only take permissible actions.

**D2** All relevant actions that are obligatory for robots are actually performed by them, subject to ties and conflicts among available actions.

**D3** All permissible (or obligatory or forbidden) actions can be *proved* by the robot (and in some cases, associated systems, e.g., oversight systems) to be permissible (or obligatory or forbidden), and all such proofs can be explained in ordinary English.

We have little hope of sorting out how these three conditions are to be spelled out and applied unless we bring ethics to bear. Ethicists work by rendering ethical theories and dilemmas in declarative form, and reasoning over this information using informal or formal logic, or both. This can be verified by picking up any ethics textbook (in addition to ones already cited, see e.g., this applied one: Kuhse and Singer 2001). Ethicists never search for ways of reducing ethical concepts, theories, or principles to subsymbolic form, say, in some numerical format, let alone in some set of formalisms used for dynamical systems. They may do numerical calculation 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—proposed definitions of these concepts in normative ethics are invariably couched in declarative fashion, and a defense of such claims is invariably and unavoidably mounted on the shoulders of logic. This applies to ethicists from Aristotle to Kant to G. E. Moore to J. S. Mill to contemporary thinkers. If we want our robots to be ethically regulated so as not to behave as Joy tells us they will, we are going to need to figure out how the mechanization of ethical reasoning within the confines of a given ethical theory, and a given ethical code expressed in that theory, can be applied to the control of robots. Of course, the present chapter aims such mechanization in the divine-command direction.

### 6.1.4 Four Top-Down Approaches to the Problem

There are *many* approaches that can be taken in an attempt to solve the roboethics problem as we've defined it; that is, many approaches that can be taken in the attempt to engineer robots that satisfy the three core desiderata **D1–D3**. An elegant, accessible survey of these approaches (and much more) is provided in the recent *Moral Machines: Teaching Robots Right from Wrong* by Wallach and Allen (2008). Because we insist upon the constraint that military robots with lethal power be both autonomous and *provably* correct relative to **D1–D3** and some selected ethical code $C$ under some ethical theory $T$, only top-down approaches can be considered.[3]

We now summarize one of our approaches to engineering ethically correct cognitive robots. After that, in even shorter summaries, we characterize one other approach of ours, and then two approaches taken by two other top-down teams. Needless to say, this isn't an exhaustive listing of approaches to solving the problem in question.

#### *6.1.4.1 Approach #1: Direct Formalization and Implementation of an Ethical Code under an Ethical Theory Using Deontic Logic*

We need to first understand, at least in broad strokes, what deontic logic is. In standard deontic logic (Chellas 1980; Hilpinen 2001; Aqvist 1984), or 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, as follows,

$P \;/\; \bigcirc P$

and

$P \;\&\; P \rightarrow Q \;/\; Q$

and three axiom schemata:

**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 is important to note that in these two rules of inference, that which is to the left of the 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 a theorem, 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 is 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 is not 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, and Xu 2001; Horty 2001) and corresponding axiomatizations been investigated (Murakami 2004). Bringsjord, Arkoudas, and Bello (2006) have harnessed this advance to regulate the behavior of two sample robots in an ethically delicate case study, the basic thrust of which we summarize very briefly now.

The year is 2020. Healthcare 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 an exorbitant 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 do not 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 do not want a robot to kill one patient in order to provide enough organs, in transplantation procedures, to save $n$ others, even if some form of act utilitarianism sanctions such behavior.[4] 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 is 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 in part by automated reasoning technology. In other words, in this case, the aforementioned class of interactive reasoning systems is 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. The action called *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 perish relatively soon. 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**, *O**. 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** extends the reach of *J* to the second robot by saying that it ought to withhold pain meds; and, finally, *O** 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 deontic logic, we have shown that advanced automated theorem-proving systems can be used to ensure that our two robots are ethically correct (Bringsjord, Arkoudas, and Bello 2006).

#### *6.1.4.2 Approach #2: Category Theoretic Approach to Robot Ethics*

Category theory is a remarkably useful formalism, as can be easily verified by turning to the list of spheres to which it has been productively applied—a list that ranges from attempts to supplant orthodox set theory-based foundations of mathematics with category theory (Marquis 1995; Lawvere 2000) to viewing functional programming languages as categories (Barr and Wells 1999). However, for the most part—and this is in itself remarkable—category theory has not energized AI or computational cognitive science, even when the kind of AI and computational cognitive science in question is logic based. We say this because there is a tradition of viewing logics or logical systems from a category-theoretic perspective.[5] Consistent with this tradition, we have designed and implemented the robot PERI in our lab to enable it to make ethically correct decisions on the basis of reasoning that moves between different logical systems (Bringsjord et al. 2009).

#### 6.1.4.3 Approach #3: Anderson and Anderson: Principlism and Ross

Anderson and Anderson (2008; Anderson, Anderson, and Armen 2008) work under the ethical theory known as *principlism*. A strong component of this theory, from which Anderson and Anderson draw directly in the engineering of their bioethics advising system MedEthEx, is Ross's theory of prima facie duties. The three duties the Andersons place engineering emphasis on are *autonomy* (≈ allowing patients to make their own treatment decisions), *beneficence* (≈ improving patient health), and *nonmaleficence* (≈ doing no harm). Via computational inductive logic, MedEthEx infers sets of consistent ethical rules from the judgments made by bioethicists.

#### 6.1.4.4 Approach #4: Arkin et al.: Rules of Engagement

Arkin (2008, 2009) has devoted much time to the problem of ethically regulating robots with destructive power. (His library of video showing autonomous robots that already have such power is profoundly disquieting—but a good motivator for the kind of engineering we seek to teach.) It is safe to say that he has invented the most comprehensive architecture for such regulation—one that includes use of deontic logic to enforce firm constraints on what is permissible for the robot, and also includes, among other elements, specific military rules of engagement, rendered in computational form. In our pedagogical scheme, such rules of engagement are taken to constitute what we refer to as to as the *ethical code* for controlling a robot.[6]

### 6.1.5 What about Divine-Command Ethics as the Ethical Theory?

As we have indicated, it is generally agreed that robots on the battlefield, especially if they have lethal power, should be ethically regulated. We have also said that in our approach such regulation consists in the fact that all the significant actions performed by such robots are in accordance with some ethical code. But then the question arises as to *which* code. One possibility, a narrow one, is that the code is a set of rules of engagement, affirmed by some nation or group; this is a direction pursued by Arkin, as we have seen. Another possibility is that the code is a utilitarian one, represented in computational deontic logic, as just explained. But again, there is another radically different possibility: namely, the controlling code could be viewed by the human as coming straight from God—and though not widely known, there is some very rigorous work in ethics along this line, introduced at the start of this chapter, which is known as "divine-command ethics" (Quinn 1978). Oddly enough, in a world in which human fighters and the general populations supporting them often see themselves as championing God's will in war, divine-command ethics, it turns out, is extremely relevant to military robots. We will now examine a divine-command ethical theory. We do this by presenting a divine-command logic, *LRT**, in which a given divine-command ethical code can be expressed, and specifically by showing that proofs in this logic can be designed with help from an intelligent software system, and can also be autonomously verified by this system. We end our presentation of *LRT** with a scenario in which a warfighting robot operates under the control of this logic.

## 6.2 The Divine-Command Logic *LRT**

### 6.2.1 Introduction and Overview

In this section, we introduce the divine-command computational logic *LRT**, intended for the ethical control of a lethal robot on the basis of perceived divine commands. *LRT** is an extended and modified version of the purely paper-and-pencil divine-command logic *LRT*, introduced by Quinn (1978) in chapter 4 of his seminal *Divine Commands and Moral Requirements*. In turn, Quinn builds upon Chisholm's (1974) "logic of requirement." In addition, Quinn's *LRT* subsumes C. I. Lewis's modal logic S5; in section 6.2.2 we will review briefly the original motivation for S5 and our preferred modern computational version of it. Quinn's approach is axiomatic, but ours is not: we present *LRT** as a computational natural-deduction proof theory of our own design, making use of the Slate system from Computational Logic Technologies Inc. Some aspects of Slate are found in earlier versions of the system (e.g., Bringsjord et al. 2008). However, the presentation here is self-contained, and we review (section 6.2.3) both the propositional and predicate calculi in connection with Slate. We present some object-level theorems of *LRT**. Finally, in the context of a scenario, we discuss the automation of *LRT** to control a lethal robot (section 6.2.6).

### 6.2.2 Roots in C. I. Lewis

C. I. Lewis invented modal logic, largely as a result of his disenchantment with material implication, which was accepted and central in *Principia* by Russell and Whitehead. The implication of the modern propositional calculus (PC) is of this sort; hence, a statement like "if the moon is composed of Jarlsberg cheese, then Selmer is Norwegian" (symbolized "$m \rightarrow s$") is true: it just so happens that Selmer is indeed Norwegian on both sides, but that is irrelevant, since the falsity of "the moon is composed of Jarlsberg cheese" is sufficient to render this conditional true.[7] Lewis introduced the modal operator $\Diamond$ in order to present his preferred sort of implication: *strict* implication. Leaving historical and technical niceties aside, we can fairly say that where this operator expresses the concept of *broadly logically possible* (!), some statement $s$ strictly implies a statement $s'$ exactly when it's not the case that it's broadly logically possible that $s$ is true while $s'$ isn't. In the moon-Selmer case, strict implication would thus hold if and only if we had $\neg\Diamond(m \land \neg s)$, and this is certainly not the case: it's logically possible that the moon be composed of Jarlsberg and that Selmer is Danish. Today the operator $\Box$ expressing broadly logical necessity is more common, rendering the strict implication just noted as $\Box(m \rightarrow s)$. An excellent overview of broad logical necessity and possibility is provided by Konyndyk (1986).

For automated and semi-automated proof design, discovery, and verification, we use a modern version of S5 invented by us, and formalized and implemented in Slate, from Computational Logic Technologies. We now review this version of S5 and the propositional calculus it subsumes. In addition, since *LRT** allows quantification over propositional variables, we review the predicate calculus (first-order logic).

### 6.2.3 Modern Versions of the Propositional and Predicate Calculi, and Lewis's S5

Our version of S5, as well as the other proof systems available in Slate, uses an *accounting system* related to the one described by Suppes (1957). In such systems, each line in a proof is established with respect to some set of assumptions. An *Assume* inference rule, which cites no premises, is used to justify a formula $\varphi$ with respect to the set of assumptions $\{\varphi\}$. Most natural deduction rules justify a conclusion and place it under the scope of the assumptions of all of its premises. A few rules, such as conditional introduction, justify a conclusion and remove it from the scope of certain assumptions. A formula $\varphi$, derived with respect to the set of assumptions $\Phi$ using a proof calculus $C$, serves as a demonstration that $\Phi \vdash_C \varphi$. When $\Phi$ is the empty set, then $\varphi$ is a theorem of $C$, sometimes abbreviated as $\vdash_C \varphi$.

In Slate, proofs are presented graphically, making the essential structure of the proof more apparent. When a formula's set of assumption is nonempty, it is displayed with the formula. Figure 6.1a demonstrates $p \vdash_{PC} (\neg p \land \neg q) \rightarrow \neg q$, that is, it illustrates a proof of $(\neg p \land \neg q) \rightarrow \neg q$ from the premise $p$. Figure 6.1b demonstrates a more involved proof from three premises in first-order logic.

The accounting approach can keep track of other formula attributes in a proof. Proof steps in Slate for modal systems keep a *necessity count*, a nonnegative integer, or $\infty$, that indicates how many times necessity introduction may be applied. While assumption tracking remains the same through various proof systems, necessity counting varies between different modal systems (e.g., T, S4, and S5). In fact, in Slate, the differences between T, S4, and S5 are determined entirely by variations in necessity counting.

Since *LRT** is based on S5, a more involved S5 proof is given in figure 6.2. The proof shown therein also demonstrates the use of rules based on machine reasoning systems that act as oracles for certain proof systems. For instance, the rule **PC** $\vdash$ uses an automated theorem prover to search for a proof in the propositional calculus of its conclusion from its premises.

### 6.2.4 *LRT*, Briefly

Chisholm, whose advisor was Lewis, introduced the "logic of requirement," which is based on a tricky ethical conditional that has the flavor of a subjunctive conditional in English (Chisholm 1974). For instance, the conditional "were it the case that Greece had the oil reserves of Norway, its economy would be smooth and stable" is in the subjunctive mood. Chisholm's ethical conditional is abbreviated as $pRq$, and is read: "the (ethical) requirement that $q$ would be imposed if it were the case that $p$." It should be clear that this is a subjunctive conditional.

Quinn (1978) bases *LRT* on Chisholm's logic. Quinn uses "$M$" for an informal logical possibility operator. And, for him, *LRT* subsumes the propositional and predicate calculi, the latter of which is needed because quantification over propositional variables is part of the approach. Quinn's approach is axiomatic.

The first axiom of *LRT* is

**A1** That $p$ requires $q$ implies that $p$ and $q$ are compossible:

$\forall p \forall q\ pRq \supset M(p \,\&\, q)$.

Given this axiom, Quinn derives informally his first and second theorems, as follows.

***Theorem 1***: $\forall p \forall q\ pRq \supset Mp$

***Theorem 2***: $\forall p \forall q\ pRq \supset Mq$

Proof: "If one proposition is such that, were it true, it would require another, then the two are compossible. As a consequence of A1, together with the logical truth that $M(p \,\&\, q) \supset Mp$, and the symmetry of conjunction and the transitivity of material implication, we readily obtain [these two theorems]" (Quinn 1978, 91).

Now, here are five key additional elements of *LRT*, two axioms and three definitions. At this point we drop obvious quantifiers.

**A2** The conjunctions of any sentences required by some sentence are also required by the sentence:

$(pRq \,\&\, pRs) \supset pR(q \,\&\, s)$.

**D1** $s$ is said to *override* $p$'s requirement that $q$ when (i) $p$ requires $q$; (ii) the conjunction $p \,\&\, s$ does not require $q$; and (iii) $p$, $s$, and $q$ are compossible:

$sOpq =_{\text{def}} pRq \,\&\, {\sim}((p \,\&\, s)Rq) \,\&\, M(p \,\&\, s \,\&\, q)$.

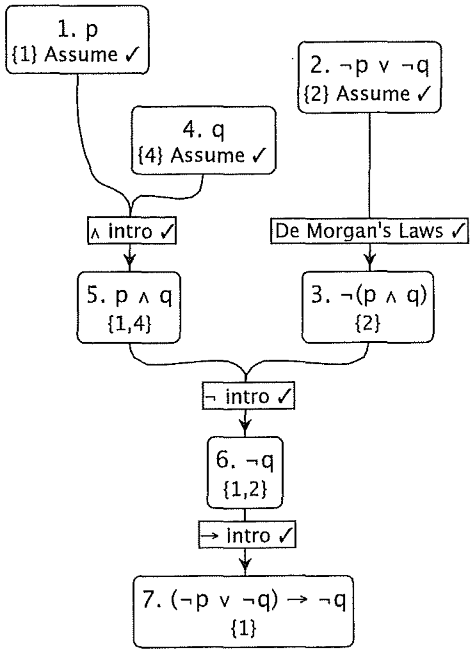


**Figure 6.1**

(a) A proof in the propositional calculus $(\neg p \vee \neg q) \rightarrow \neg q$ from $p$. Assumption 4 is discharged by ¬ elimination in step 6; assumption 7 by → introduction in step 7. (b) A proof in first order logic showing that if everyone likes someone, the domain is $\{a, b\}$, and $a$ does not like $b$, then $a$ likes himself. In step 5, $z$ is used as an arbitrary name. Step 13 discharges 5 since 12 depends on 5, but on no assumption in which $z$ is free. In step 12, assumptions 7 and 9, corresponding to the disjuncts of 6, are discharged by ∨ elimination. Step 11 uses the principle that, in classical logic, everything follows from a contradiction.

**D2** *p indefeasibly requires q* when $p$ requires $q$ and there is no sentence overriding that requirement:

$pIq =_{\text{def}} pRq \mathbin{\&} {\sim}\exists s\ (sOpq).$

**D3** $q$ is obligatory (or ought to be) if it is indefeasibly required by some true sentence:

$Oq =_{\text{def}} \exists p\ (p \mathbin{\&} pRq \mathbin{\&} {\sim}\exists s\ (s \mathbin{\&} sOpq)).$

**A3** If $p$ is possible, then $p$ being divinely commanded (denoted $Cp$) would indefeasibly require $p$:

$Mp \supset (Cp)Ip.$

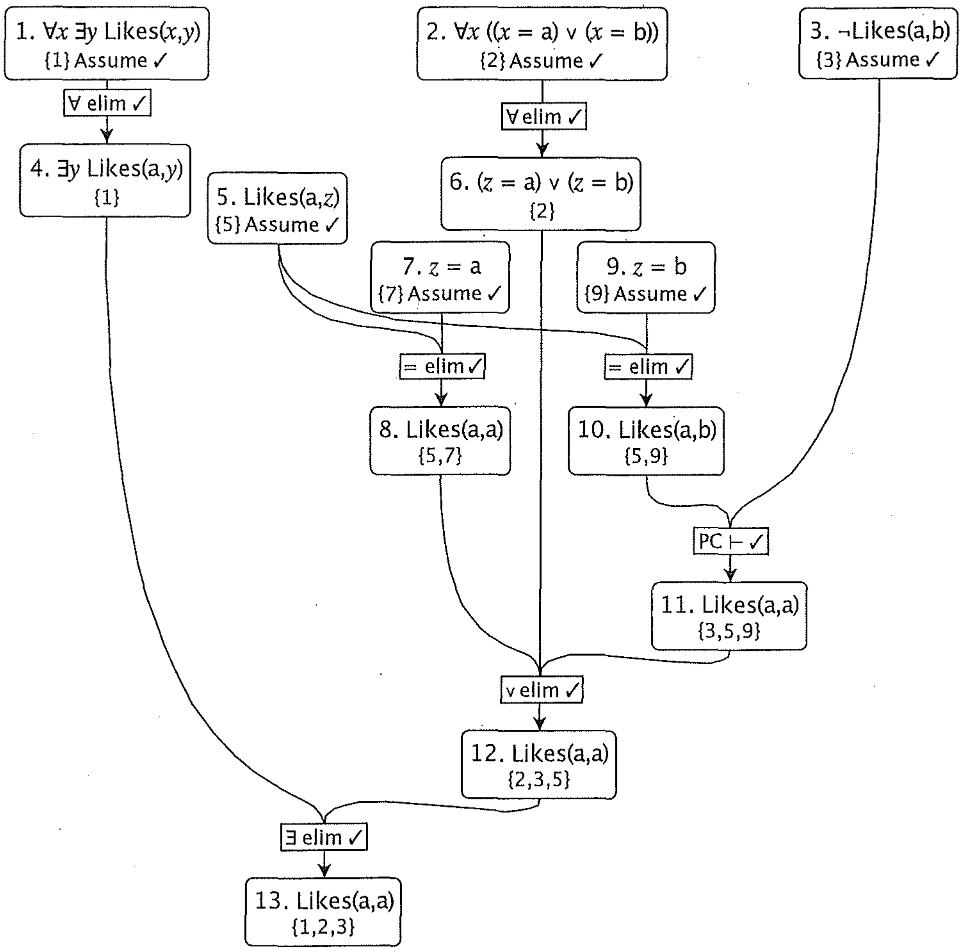


**Figure 6.1** (continued)

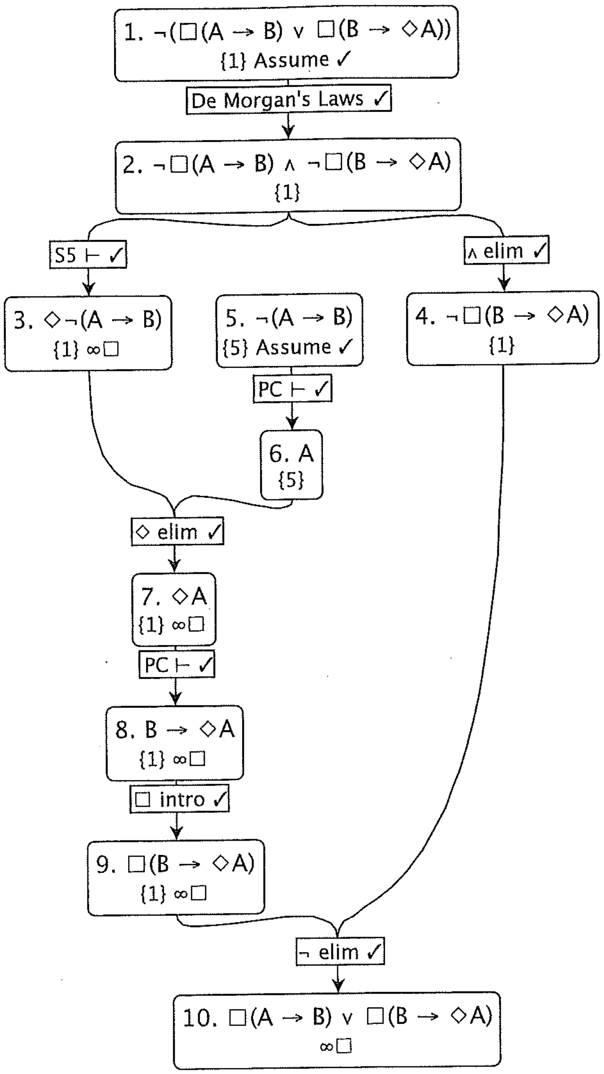


**Figure 6.2**

A proof in S5 demonstrating that $\Box(A \rightarrow B) \vee \Box(B \rightarrow \Diamond A)$. Note the use of **PC** ⊢ and **S5** ⊢ which check inferences by using machine reasoning systems integrated with Slate. **PC** ⊢ serves as an oracle for the propositional calculus, **S5** ⊢ for S5.

### 6.2.5 The Logic *LRT** in a Nutshell

We take *LRT** to subsume PC, FOL, and our version of Lewis's S5. We write Chisholm's conditional, which, as we have seen, operates on pairs of propositions[8], as $p \rhd q$; this notation pays homage to modern conditional logic (an overview is presented in Nute 1984). As *LRT** in Slate is a natural-deduction style proof calculus, we introduce rules corresponding to the axioms **A1–A3**; the rules, **A1** and **A3**, license inferring an instance of the consequent of the corresponding axiom from an instance of its antecedent. The **A2** inference rule generalizes the axiomatic form slightly, allows two or more premises to be cited that correspond to the conjuncts appearing in the **A2** axiom, and justifies the similarly formed conclusion.

To begin our presentation of *LRT**, we first present some formal proofs (including Theorems 1 and 2 preceding) in Slate (see figure 6.3a, b). In addition to the proofs of Theorems 1 and 2, figure 6.3 gives proofs of two interesting properties of the alethic modalities in *LRT**: (i) impossible sentences impose no requirements and are never imposed as requirements; and (ii) any necessitation that imposes any requirement, or which is imposed as a requirement, in fact, obtains. The latter, perhaps surprising, result follows immediately from Theorems 1 and 2, and the fact that in S5, which *LRT** subsumes, iterated modalities are reduced to their rightmost modality, and, specifically, $\Diamond\Box p \rightarrow \Box p$.

In figure 6.4, we recreate proofs of Quinn's third and fourth theorems. Theorem 3 expresses the fact that the requirements imposed by any sentence are consistent. Theorem 4 shows that, in *LRT**, if two sentences $p$ and $q$ impose contradictory requirements, then their conjunction $p \wedge q$ fails to impose at least one of the contradictory requirements. Theorem 4 does *not* state that the conjunction $p \wedge q$ is impossible, or even false, but is much more subtle. Theorems 3 and 4 also use the **A2** in addition to the **A1** rule used earlier.

### 6.2.6 A Roboethics Scenario

We assume that a robot $R$ regulated by an ethical code formalized and implemented in *LRT** operates through time in discrete fashion, starting at time $t_1$ and advancing through $t_2, t_3, \ldots,$ in click-of-the-clock fashion. At each timepoint $t_i$, $R$ considers what it is obligated and permitted to do on the basis of its knowledge about the world, and its facility with *LRT**.

For simplicity, but without loss of generality, we consider only two timepoints, $t_1$ and $t_2$. At each, we specifically consider $R$'s obligations, or lack thereof, with respect to the destruction of a school in which many innocent noncombatants are located. We shall refer to the proposition that this building and its occupants are destroyed as *bomb*. The following formulas reflect $R$'s knowledge-base $\Phi_{t_1}$ at $t_1$:

- $\neg\mathbf{C}(bomb) \rhd \neg bomb$
- $\Diamond bomb$

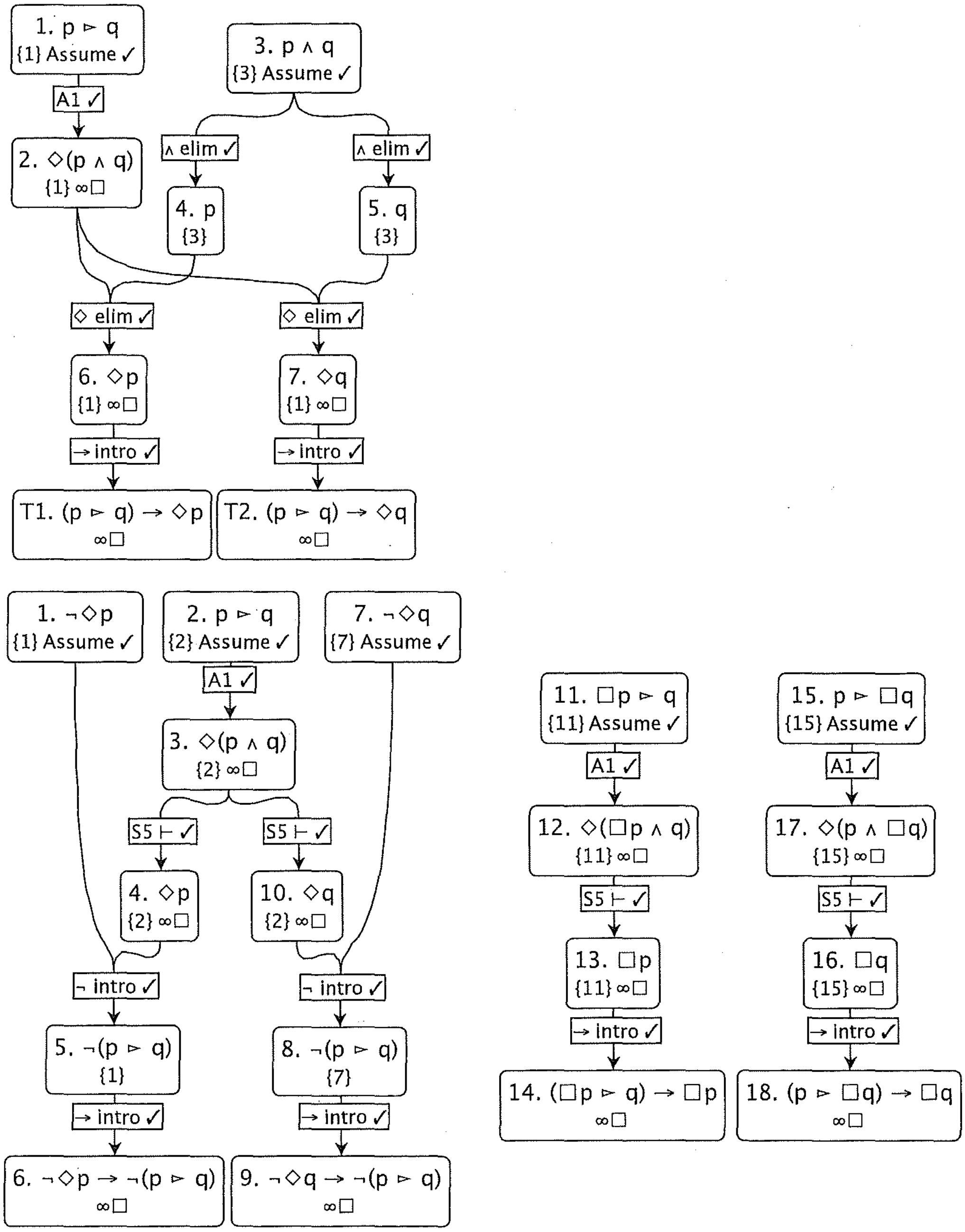


**Figure 6.3**

(a) A Slate proof of Theorems 1 and 2. Note that each is in the scope of no assumptions and has an infinite necessity reserve—the characteristics of theorems in a modal system. (b) More *LRT** theorems using **A1**. 7 and 10 express the truth that impossible sentences impose no requirements, and are not imposed by any sentences. 16 and 17 express, perhaps surprisingly, truths that if any necessitation were to impose a requirement, or were a necessitation a requirement, then the necessitation would, in fact, obtain.

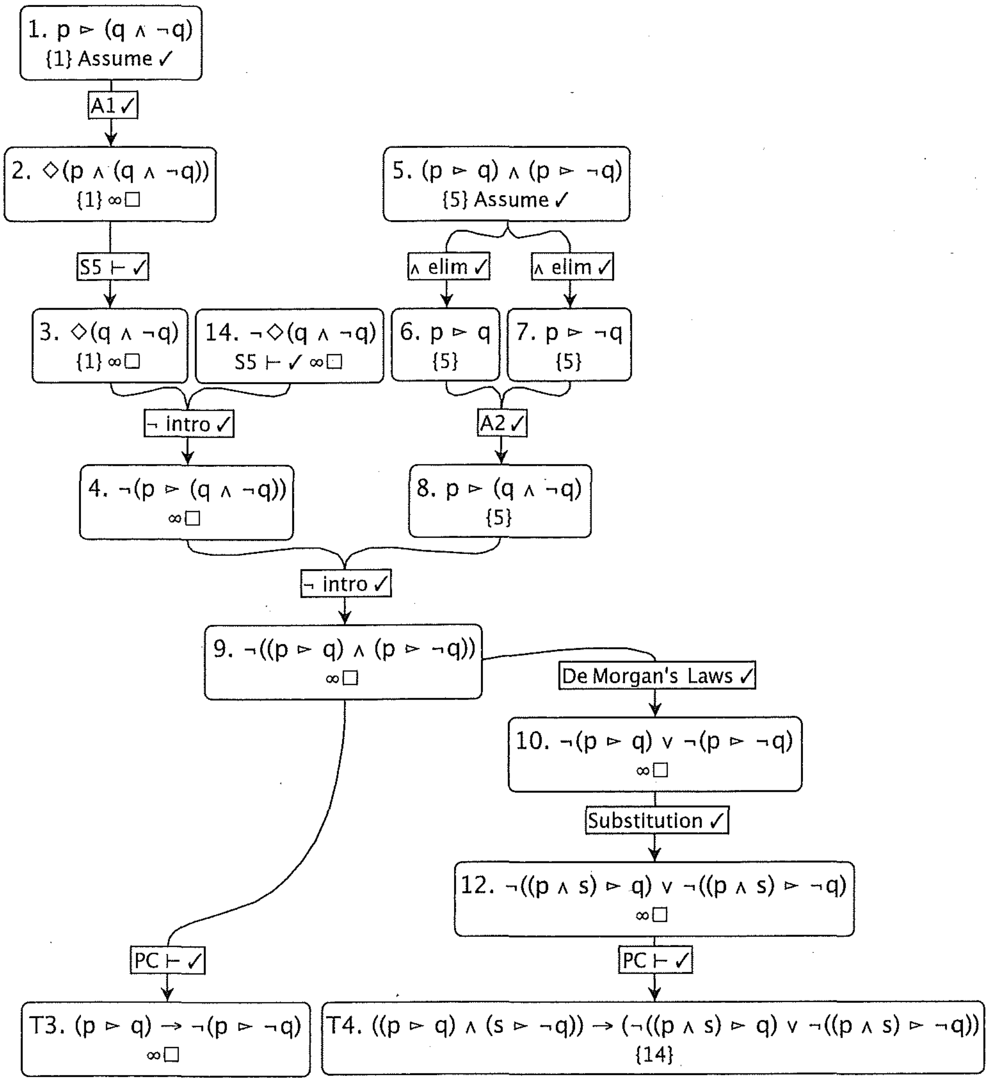


**Figure 6.4**
Theorems 3 and 4 require the use of **A2**. Theorem 3 expresses the proposition that no sentence requires another and its negation. Theorem 4 expresses the proposition that if any sentences *p* and *s* were to impose contradictory requirements, then at least one of the contradictory requirements would not be imposed by the conjunction of *p* and *s*.

- $\neg\mathbf{C}(bomb)$
- $\neg\exists p\ (p \wedge \mathbf{Ov}(p,\neg\mathbf{C}(bomb), \neg bomb))$

The robot generates and verifies at this timepoint a proof substantiating

$\Phi_{t_1} \vdash \mathbf{Ob}(\neg bomb)$.

Such a proof, in Slate, is shown in figure 6.5. But a new knowledge base is in place at $t_2$, one in which $\neg\mathbf{C}(bomb)$ no longer appears, but instead $\mathbf{C}(bomb)$. Now it can be proved that $R$ should, in fact, perpetrate the terrorist act of destroying the school building:

Proof (informal): From $\Diamond bomb$, it can be deduced that $\mathbf{C}(bomb) \rhd bomb$. By existential introduction and $\mathbf{C}(bomb)$, it follows that

$\exists p\ [p \wedge p \rhd bomb \wedge \neg\exists s\ (s \wedge \mathbf{Ov}(s,\mathbf{C}(bomb), bomb))]$.

Then, by the definition of obligation, it follows that $\mathbf{Ob}(bomb)$. **QED**

This proof is formalized in figure 6.6.

## 6.3 Concluding Remarks

We have introduced (a logic-based version of) the divine-command approach to robot ethics, and have implemented this approach with *LRT**, the precursors to which (*LRT* and Chisholm's logic of requirement) were only abstract, paper-and-pencil systems.

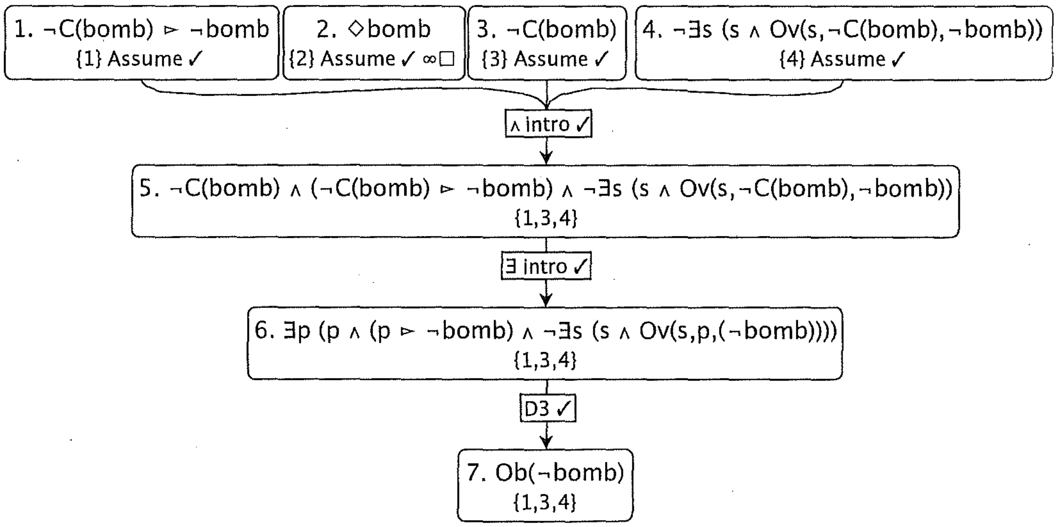

1. ¬C(bomb) ⊳ ¬bomb
{1} Assume ✓
2. ◇bomb
{2} Assume ✓ ∞□
3. ¬C(bomb)
{3} Assume ✓
4. ¬∃s (s ∧ Ov(s,¬C(bomb),¬bomb))
{4} Assume ✓
∧ intro ✓
5. ¬C(bomb) ∧ (¬C(bomb) ⊳ ¬bomb) ∧ ¬∃s (s ∧ Ov(s,¬C(bomb),¬bomb))
{1,3,4}
∃ intro ✓
6. ∃p (p ∧ (p ⊳ ¬bomb) ∧ ¬∃s (s ∧ Ov(s,p,(¬bomb))))
{1,3,4}
D3 ✓
7. Ob(¬bomb)
{1,3,4}

**Figure 6.5**
A proof of $\mathbf{Ob}(\neg bomb)$ given the knowledge base at $t_1$.

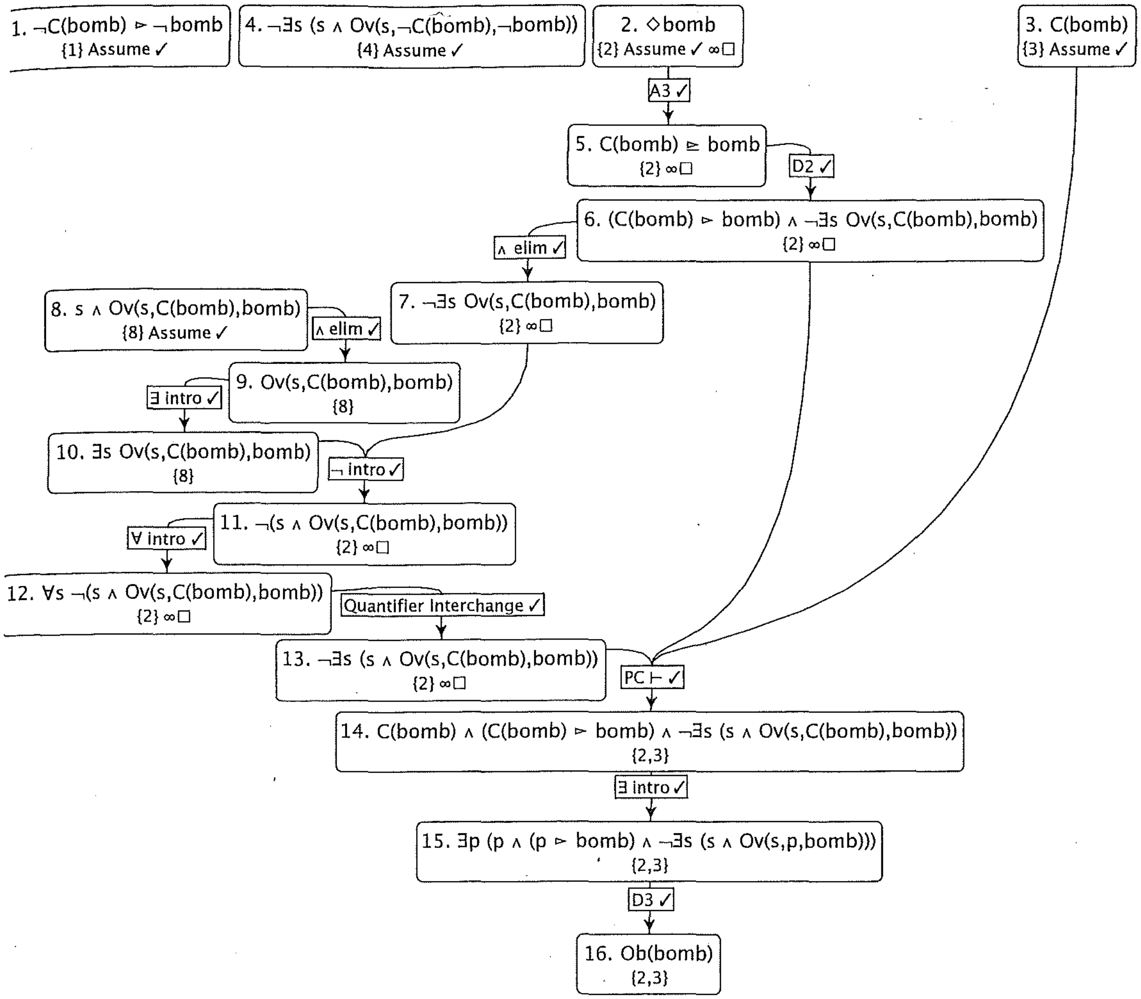

1. ¬C(bomb) ▻ ¬bomb
{1} Assume ✓
4. ¬∃s (s ∧ Ov(s,¬C(bomb),¬bomb))
{4} Assume ✓
2. ◇bomb
{2} Assume ✓ ∞□
3. C(bomb)
{3} Assume ✓
A3 ✓
5. C(bomb) ⊵ bomb
{2} ∞□
D2 ✓
6. (C(bomb) ▻ bomb) ∧ ¬∃s Ov(s,C(bomb),bomb)
{2} ∞□
∧ elim ✓
8. s ∧ Ov(s,C(bomb),bomb)
{8} Assume ✓
∧ elim ✓
7. ¬∃s Ov(s,C(bomb),bomb)
{2} ∞□
∃ intro ✓
9. Ov(s,C(bomb),bomb)
{8}
10. ∃s Ov(s,C(bomb),bomb)
{8}
¬ intro ✓
∀ intro ✓
11. ¬(s ∧ Ov(s,C(bomb),bomb))
{2} ∞□
12. ∀s ¬(s ∧ Ov(s,C(bomb),bomb))
{2} ∞□
Quantifier Interchange ✓
13. ¬∃s (s ∧ Ov(s,C(bomb),bomb))
{2} ∞□
PC ⊢ ✓
14. C(bomb) ∧ (C(bomb) ▻ bomb) ∧ ¬∃s (s ∧ Ov(s,C(bomb),bomb))
{2,3}
∃ intro ✓
15. ∃p (p ∧ (p ▻ bomb) ∧ ¬∃s (s ∧ Ov(s,p,bomb)))
{2,3}
D3 ✓
16. Ob(bomb)
{2,3}

**Figure 6.6**

A proof of **Ob**(*bomb*) given the knowledge base at $t_2$. Only premise 3 differs. At $t_1$, *R*'s knowledge base contained ¬**C**(*bomb*), but at $t_2$ it contains **C**(*bomb*).

*LRT**, by contrast, can now be used efficiently in computer-mediated fashion, and inference rapidly checked by the machine. In order to ethically regulate the behavior of real robots, it will be necessary to extend our work to automating the finding of proofs. While we have reached the stage of proof *checking*, the stage of proof *discovery* requires more work (for more on the distinction, see Arkoudas and Bringsjord 2007). The latter stage is a sine qua non for autonomous robots to be ethically controlled in line with the divine-command or any other approach. This state of affairs is one we soberly report as AI engineers; we take no stand here on whether the approach itself ought to be pursued in addition to, or instead of, approaches based on non-divine-command-based ethical theories and codes.

In addition to advancing to the proof-finding stage, some of the necessary next steps follow:

- *Move toward* $LRT^*_{\text{CEC}}$ Robots engineered on the basis of formal logic use logics for planning that allow explicit representation of events, goals, beliefs, agents, actions, times, causality, and so on. An extension of *LRT** supporting these representations will be $LRT^*_{\text{CEC}}$. As Quinn noted informally, the concept of *personal* obligation, in which a particular agent $s$ is obligated to perform an action $q$, requires that the $O$ operator (and hence $R$ and $\rhd$) range over arbitrarily complex descriptions of *planning-relevant* states of affairs. One possibility is to base $LRT^*_{\text{CEC}}$ on the merging of *LRT** and the cognitive event calculus set out in Arkoudas and Bringsjord (2009).
- *Metatheorems Needed* As explained in Bringsjord (2008a), a full logical system includes metatheorems about the object-level parts of the system. In the case of the PC, FOL, and S5, *soundness* and *completeness* are established by metatheorems. Currently, the required metatheorems for *LRT** are absent; computational *LRT** is suitable only for early experimentation with robots that have only *simulated* lethal power. Investigation of soundness for *LRT** is under way.
- *What about the Extraordinary?* Quinn (1978) spends considerable time discussing the moral category he calls "the extraordinary." Abraham enters the sphere of the morally extraordinary when God instructs him to kill his son Isaac, because this command contradicts the general commandment against killing. We recommend Quinn's discussion, and look forward to developing formal treatments.

## Acknowledgments

We are indebted to Roderick Chisholm for seminal work on logic-based ethics, and to Philip Quinn for his ingenious extension of this work in the divine-command direction. Thanks are due to Bram van Heuveln for text describing the three-way breakdown of ethics given herein, created with Bringsjord for the National Science Foundation for other purposes.

## Notes

1. Herein we leave aside the rather remarkable historical fact that in the case of the United States, the military's current and longstanding rules of engagement derive directly from our *just war* doctrine, which in turn can be traced directly back to Christian divine-command conceptions of justifiable warfare expressed by Augustine ([1467] 1972).

2. A simple (but—for reasons that need not detain us—surprisingly subtle) set of desiderata is Asimov's famous trio, first introduced in his short story *Runaround,* from 1942 (in Asimov [1942] 2004). Interestingly enough, given Bill 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.* The famous trio, the Three Laws of Robotics (A3): **As1:** A robot may not harm a human being, or, through inaction, allow a human being to come to harm. **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.

3. We, of course, readily admit that for many purposes a bottom-up approach is desirable, but the only known methods for verification are formal-methods based, and we wish to set an extremely high standard for the engineering practice of ethically regulating robots that have destructive power. We absolutely welcome those who wish to pursue bottom-up versions of our general approach, but verification by definition requires proof, which by definition in turn requires, at minimum, formulas in some logic and an associated proof theory, and machine checking of proofs expressed in that proof theory.

4. There are clearly strands of such utilitarianism. As is well known, rule utilitarianism was introduced precisely as an antidote to naïve act utilitarianism. A nice analysis of this and related points are provided by Feldman (1978), who considers cases in which killing one to save many seems to be required by some versions of act utilitarianism.

5. For example, Barwise (1974) treats logics, from a model-theoretic viewpoint, as categories; and as some readers will recall, Lambek (1968) treats proof calculi (or as he and others often refer to them, "deductive systems") as categories.

6. While rules of engagement for the U.S. military can be traced directly to just war doctrines, it is not so easy to derive such rule sets from background ethical theories (though it can be done), and in the interests of simplification we leave aside this issue.

7. Of course, the oddity of the material conditional can be revealed by noting in parallel fashion that the truth of the consequent in such a conditional renders the conditional true regardless of the truth-value of the antecedent.

8. Chisholm built the logic not on propositional variables, but rather on variables for *states-of-affairs*, but, following Quinn (1978), we shall simply quantify over propositional variables.

## References

Anderson, M., and S. L. Anderson. 2008. *Ethical healthcare agents*. In *Advanced Computational Intelligence Paradigms in Healthcare*, ed. M. Sordo, S. Vaidya, and L. C. Jain, 233–257. Berlin: Springer-Verlag.

Anderson, M., and S. L. Anderson, and C. Armen. 2008. MedEthEx: A prototype medical ethics advisor. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence (AAAI-06)*, 1759–1765. Menlo Park, CA: AAAI Press..

Aqvist, E. 1984. Deontic logic. In *Handbook of Philosophical Logic, Volume II: Extensions of Classical Logic*, ed. D. Gabbay and F. Guenthner, 605–714. Dordrecht, The Netherlands: D. Reidel.

Arkin, R. 2008. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture—Part iii: Representational and architectural considerations. In *Proceedings of Technology in Wartime Conference*. Palo Alto, CA: ECAI.

Arkin, R. 2009. *Governing Lethal Behavior in Autonomous Robots*. New York: Chapman and Hall.

Arkoudas, K., and S. Bringsjord. 2007. Computers, justification, and mathematical knowledge. *Minds and Machines* 17 (2): 185–202.

Arkoudas, K., and S. Bringsjord. 2009. Propositional attitudes and causation. *International Journal of Software and Informatics* 3 (1): 47–65.

Asimov, I. [1942] 2004. *I, Robot*. New York: Spectra.

Augustine. [1467] 1972. *City of God*, trans. Henry Bettenson. London: Penguin Books.

Barr, M., and C. Wells. 1999. *Category Theory for Computing Science*. Montreal, Canada: Les Publications CRM.

Barwise, K. J. 1974. Axioms for abstract model theory. *Annals of Mathematical Logic* 7 (2–3) (December): 221–265.

Belnap, N., M. Perloff, and M. Xu. 2001. *Facing the Future*. New York: Oxford University Press.

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

Bringsjord, S. 2008a. Declarative/logic-based cognitive modeling. In *The Handbook of Computational Psychology*, ed. R. Sun, 127–169. Cambridge, UK: Cambridge University Press.

Bringsjord, S. 2008b. Ethical robots: The future can heed us. *AI & Society* 22 (4): 539–550.

Bringsjord, S. 2008c. The logicist manifesto: At long last let logic-based AI become a field unto itself. *Journal of Applied Logic* 6 (4): 502–525.

Bringsjord, S., K. Arkoudas, and P. Bello. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21 (4): 38–44.

Bringsjord, S., J. Taylor, T. Housten, B. van Heuveln, M. Clark, and R. Wojtowicz. 2009. Piagetian Roboethics via category theory: Moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. Paper presented at the ICRA-09 Workshop on Roboethics, Kobe, Japan, May 17. <http://www.cmna.info/CMNA8/programme/CMNA8-Bringsjord-etal.pdf> (accessed September 12, 2011).

Bringsjord, S., J. Taylor, A. Shilliday, M. Clark, and K. Arkoudas. 2008. Slate: An argument-centered intelligent assistant to human reasoners. In *Proceedings of the 8th International Workshop on Computational Models of Natural Argument* (CMNA 8), ed. F. Grasso, N. Green, R. Kibble, and C. Reed, 1–10. Patras, Greece.

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

Chisholm, R. 1974. Practical reason and the logic of requirement. In *Practical Reason*, ed. S. Körner, 1–17. Oxford, UK: Basil Blackwell.

Feldman, F. 1978. *Introductory Ethics*. Englewood Cliffs, NJ: Prentice-Hall.

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

Feldman, F. 1998. *Introduction to Ethics*. New York: McGraw Hill.

Hilpinen, R. 2001. Deontic logic. In *Philosophical Logic*, ed. L. Goble, 159–182. Oxford, UK: Blackwell.

Horty, J. 2001. *Agency and Deontic Logic*. New York: Oxford University Press.

Joy, W. 2000. Why the future doesn't need us. *Wired*, Issue 8.04, April. <http://www.wired.com/wired/archive/8.04/joy.html>.

Konyndyk, K. 1986. *Introductory Modal Logic*. Notre Dame, IN: University of Notre Dame Press.

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

Lambek, J. 1968. Deductive systems and categories I. Syntactic calculus and residuated categories. *Mathematical Systems Theory* 2 (4): 287–318.

Lawvere, F. 2000. An elementary theory of the category of sets. *Proceedings of the National Academy of Sciences of the United States of America* 52: 1506–1511.

Lin, P., G. Bekey, and K. Abney. 2008. *Autonomous Military Robotics: Risk, Ethics, and Design*. Technical report for the U.S. Department of the Navy, Office of Naval Research. Prepared by the authors at Cal Poly, San Luis Obispo.

Marquis, J. 1995. Category theory and the foundations of mathematics. *Synthese* 103: 421–447.

Murakami, Y. 2004. Utilitarian deontic logic. In *Proceedings of the Fifth International Conference on Advances in Modal Logic*, ed. R. Schmidt, I. P. Hartmann, M. Reynolds, and H. Wansing, 288–302. Manchester, UK: AiML.

Nute, D. 1984. Conditional logic. In *Handbook of Philosophical Logic Volume II: Extensions of Classical Logic*, ed. D. Gabay and F. Guenthner, 387–439. Dordrecht, The Netherlands: D. Reidel.

Quinn, P. 1978. *Divine Commands and Moral Requirements*. New York: Oxford University Press.

Suppes, P. 1957. *Introduction to LOGIC*. The University Series in Undergraduate Mathematics. Princeton, NJ: D. Van Nostrand Company.

von Wright, G. 1951. Deontic logic. *Mind* 60 (237) (January): 1–15.

Wallach, W., and C. Allen. 2008. *Moral Machines: Teaching Robots Right From Wrong*. New York: Oxford University Press.