Introducing Divine-Command Robot Ethics*

Selmer Bringsjord & Joshua Taylor
Department of Computer Science
Department of Cognitive Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu • tayloj@rpi.edu

062310

Abstract

Perhaps it is generally agreed that robots on the battlefield, especially if they have lethal power, should be ethically regulated. But in what does such regulation consist? Presumably in the fact that all the significant actions performed by such a robot 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. Another possibility is that the code is a utilitarian one represented in computational deontic logic, as explained elsewhere by Bringsjord and colleagues. Another possibility is likewise based on computational logic, but with a logic that captures some other mainstream ethical theory (e.g., Kantian deontology, or Ross' "right mix" direction). But there is another radically different possibility that hitherto has not arrived on the scene: viz., the controlling code could be viewed by the human as coming straight from God. There is some very rigorous work in ethics along this line, which is known as *divine-command ethics*. In a world in which human fighters and the general populations supporting them often see themselves as indeed championing God's will in war, divine-command ethics is quite relevant to military robots. This chapter introduces divine-command ethics in the form of the computational logic $\mathcal{LRT}^\star$, intended to eventually be suitable for regulating a real-world warfighting robot.

Contents

*We are indebted to Roderick Chisholm for seminal work on logic-based ethics, and to Phil 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.

1 Introduction

Perhaps it's generally agreed that robots on the battlefield, especially if they have 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 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 approach, described below, has been taken by Arkin (2009, 2008).[1] Another possibility is that the code is a utilitarian one represented in computational deontic logic, as explained for instance by Bringsjord et al. (2006), and summarized below. Another possibility is likewise based on computational logic, but using a logic that captures some other mainstream ethical theory (e.g. Kantian deontology, or Ross' "right mix" direction); this possibility has been rigorously pursued by Anderson & Anderson (2008, 2006). But there is a radically different possibility that hitherto hasn't arrived on the scene: viz., the controlling moral code could be viewed by the human as coming straight from God. There is some very rigorous work in ethics along this line, which is 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 in general a monotheistic scheme, it seems more than peculiar that heretofore robot ethics (or as we prefer to say, *roboethics*) has been bereft of the systematic study of such ethics on the basis of monotheistic conceptions of what is right and wrong, morally speaking. This chapter introduces divine-command ethics in the form of the computational logic $\mathcal{LRT}^\star$, 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 2008*c*).

The paper is structured as follows. We first set out the general context of roboethics in a military setting (§2), and simply point out that the divine-command approach has been absent from that context. We then introduce the divine-command computational logic $\mathcal{LRT}^\star$ (§3), concluding this section with a robot scenario in which the robot in question is constrained by dynamic use of the logic. We end (§4) with some remarks about next steps in the divine-command roboethics program.

2 The Context for Divine-Command Roboethics

2.1 Tripartite Partition of Ethics

There are several branches of ethics. One standard tripartite breakdown is as follows. In this scheme, the second and third branches directly connect to our roboethics R&D; we discuss the connection immediately after summarizing the trio. But first a note about books and resources from which our breakdown can be derived.

There are of course many textbooks that provide excellent overviews. One text we trust greatly, and which we have used in teaching "straight" introduction-to-ethics courses and modules, is (Feldman 1978), a true, enduring classic (not to be confused with the more recent (Feldman 1998), which is an anthology of readings). One nice feature of this book is that it conforms with arguably the most sophisticated published presentation of utilitarianism from the standpoint of the semantics of deontic logic: (Feldman 1986). Since much of our prior R&D has been based on deontic logic (e.g., see Bringsjord et al. 2006), this second book is part of the foundation for our work in roboethics.

[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).

Now to the three branches of ethics, with apologies in advance to those readers who are ethicists, or close.

Meta-ethics *Meta-ethics* 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 more like matters of opinion? Who or what determines whether something is good or bad? Is there a divine being who stipulates what is right or wrong? Is there some kind of Platonic realm of ethics that provides truth-values to ethical claims independently of what anyone thinks? Or is ethics merely "in the head," and if so, how can any one's moral outlook be seen as "better" than any other's?

As AI engineers trying to bestow ethical qualities to robots (in a manner soon to be explained), we are automatically confronted with various of these meta-ethical issues, especially given the power we have to determine a robot's "sense" of right and wrong. Is the robot's sense of right and wrong just an arbitrary choice of the programmer? Or are there objective guidelines to determine whether the moral outlook of one robot is better than any of that of any other robot or, for that matter, that of a human? Are we playing God? (Perhaps a somewhat ironic question in the present context, devoted as it is to presenting a form of roboethics based on God's commands.) By 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 Whereas meta-ethics is highly abstract and theoretical, the field of *applied ethics* is much more practical and specific. Rather than trying to divine where ethical guidelines come from, applied ethics *starts* with a certain set of moral guides, but then applies them to specific domains to try and answer specific moral dilemmas that arise in that particular area of interest. 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 the fact 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? Also, while many of these issues overlap with other fields of engineering, the potential of robots to become—courtesy of the very engineering we are in the business of carrying out—ethical agents themselves raises an additional set of moral questions: Do such robots have any rights and responsibilities? Both sets of these kinds of considerations can be seen as part of roboethics.

Normative Ethics In the field of *normative ethics*, or *moral theory*, one compares and contrasts different ways by which one can define the concepts *obligatory*, *forbidden*, *permissible*, and *supererogatory*. (Sometimes synonymous terms are used; e.g., *wrong* for *forbidden*.) Normative ethics investigates which actions we ought to, or ought not to, perform, and why. A fundamental distinction in normative ethics is that between *consequentialist* views, which state that actions are good or bad depending on their outcomes, and *non-consequentialist* views, which instead consider the intent behind an action, or the inherent duties, rights, and responsibilities that may be involved, independent of what the actual outcome of the action turns out to be. Well-known consequentialist views are egoism, altruism, and utilitarianism, while the best known non-consequentialist view is

probably Kant's deontological theory of moral behavior, the kernel of which is the principle that people should never be treated as a means to an end.

2.2 Where Our Work Falls

In our roboethics work we are for the most part working within normative ethics, and in two important ways. First, given any particular normative theory $\mathcal{T}$, we take on the burden of needing to find a way to engineer a robot with that particular outlook by deriving and specializing from $\mathcal{T}$ a particular ethical code $\mathcal{C}$ that fits the robot's environment. In particular, we accept the burden of providing in advance a *guarantee* that a lethal robot does indeed adhere to a particular moral code $\mathcal{C}$ under some theory $\mathcal{T}$. Second, once robots are infused with a particular ethical code $\mathcal{C}$, the robots can be placed under different conditions to see how the different codes play out. Strengths and weaknesses of the different ethical codes can thus be observed and empirically studied; this may well inform the field of normative ethics itself.

Our work can be viewed as falling "between" the areas of meta-ethics and applied ethics. Like meta-ethics, normative ethics does not look at specific moral dilemmas as they occur in the real world, but instead keeps the different theories general and thus applicable to any domain. However, like applied ethics, normative ethics doesn't ask for the deep metaphysical status of any of these theories, but rather takes these theories as they are, and considers the different outcomes of these theories as they are being applied.

2.3 The Importance of Robot Ethics

As is well-known, 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 (2008*b*) rejects this fatalism, if we assume that robots in the future will have more and more autonomy, and more and more lethal power, it certainly seems reasonable to at least 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. For example, the importance of such engineering is now openly and aggressively affirmed by the Defense community. For example, a recent extensive and enlightening summary of the overall landscape is provided by Lin et al. (2008). In this thorough summary for the Office of Naval Research, the possibility and need of creating ethical robots is analyzed. The recommended goal of Lin et al. (2008) is not to make fully ethical machines, but simply machines that perform better than humans in isolated cases. Lin et al. (2008) 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 the robots discussed in the Lin et al. (2008) report are mostly controlled by remote human warfighters, the Department of Defense's Unmanned Systems Integrated Roadmap supports the desire for increasing the autonomy of these robots. We thus frankly view the problem as follows: Gradually, because of economic and social pressures that will be impossible to suppress, autonomous warfighting robots with lethal power will be deployed in all theaters of war. These pressures are already clearly in play. For example, in a world where expenditures for defense and social programs increasingly outstrip revenues from taxation, cost-cutting via taking expensive humans out of the loop will soon enough prove irresistible. So while humans are still firmly in the so-called "kill chain" today, their gradual removal from this chain in favor of inexpensive and expendable robots is as inevitable as the rising of the sun. Even if our pessimism is incorrect, only thinkers in the grip of pollyana views of the future would resist our call to at the very least plan for

Figure 1: One of the many partially autonomous military ground robots of today.

the *possibility* that the black future we envision may unfold—and such prudent planning is all that we need to motivate the roboethical engineering that we are in general calling for.

2.4 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 and/or formal logic. This can be verified by picking up any ethics textbook (in addition to ones already cited, see e.g., this applied one: Kuhse & Singer 2001). Ethicists never search for ways of reducing ethical concepts, theories, principles to sub-symbolic 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

[2] A simple (but—for reasons that need not detain us—surprisingly subtle) set of desiderata are Asimov's famous trio, first introduced in his short story "Runaround," from 1942. (You can find the story in (Asimov 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 ($\mathcal{A}3$): **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.

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 is 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.

2.5 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 & 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 $\mathcal{C}$ under some ethical theory $\mathcal{T}$, 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.

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 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.,

$$\frac{P}{\bigcirc P} \qquad \text{and} \qquad \frac{P \quad P \rightarrow Q}{Q}$$

and three axiom schemata:

A1 All tautologous well-formed formulas.

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

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

[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 require 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.

It is 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 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 & Xu 2001, Horty 2001) and corresponding axiomatizations been investigated (Murakami 2004). Bringsjord et al. (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. 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 an exorbitant pain medication. Of paramount importance, obviously, is that neither robot perform an action that is morally wrong according to the ethical code $\mathcal{C}$ 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 $\mathcal{C}' \neq \mathcal{C}$, 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 strand 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., $\mathcal{C}$ 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

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

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 deontic logic, we have shown that advanced automated theorem proving systems can be used to ensure that our two robots are ethically correct (Bringsjord et al. 2006).

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 & 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).

Approach #3: Anderson & Anderson: Principlism and Ross

Anderson & Anderson (2008, 2006) work under the ethical theory known as *principlism*. A strong component of this theory, from which Anderson & 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* ($\approx$ allow patients to make their own treatment decisions), *beneficence* ($\approx$ improve patient health), and *nonmaleficence* ($\approx$ do no harm). Via computational inductive logic, MedEthEx infers sets of consistent ethical rules from the judgments made by bioethicists.

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

Arkin (2009, 2008) 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

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

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]

2.6 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 a robot 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 explained immediately above. But there is another radically different possibility: viz., 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, 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 now introduce a divine-command ethical theory. We do this by presenting a divine-command logic, $\mathcal{LRT}^\star$, 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 $\mathcal{LRT}^\star$ with a scenario in which a warfighting robot operates under the control of this logic.

3 The Divine-Command Logic $\mathcal{LRT}^\star$

3.1 Introduction and Overview of the Section

In this section we introduce the divine-command computational logic $\mathcal{LRT}^\star$, intended for the ethical control of a lethal robot on the basis of perceived divine commands. To our knowledge, this is the first such logic of its kind. In keeping with what we said at the outset, that the advent of such a logic, in a world like the one we find ourselves in, takes place now, long after non-divinity-based approaches to the ethical control of autonomous robots have appeared, is something we find remarkable.

$\mathcal{LRT}^\star$ is an extended and modified version of the purely paper-and-pencil divine-command logic *LRT*, introduced by Quinn (1978) in Chapter IV of his seminal *Divine Commands and Moral Requirements*. In turn, Quinn, as he makes explicit, builds, in presenting *LRT*, directly upon Chisholm's (1974) "logic of requirement." In addition, though Quinn isn't clear about the issue, his *LRT* subsumes C. I. Lewis' modal logic S5, the original motivation for which, and our preferred modern computational version of which, we briefly review (§ 3.2). While as we indicate (§ 3.4) Quinn's approach is an axiomatic one, ours is not: We present $\mathcal{LRT}^\star$ (§ 3.5) in the form of a computational natural-deduction proof theory of our own design, making use of the **X** system from Computational Logic Technologies Inc. Some aspects of **X** are found in the Slate system; see (Bringsjord et al. 2008). But the presentation here is self-contained, and indeed in section 3.3 we review for the reader both the propositional and predicate calculi in connection with **X**. We present some object-level theorems of $\mathcal{LRT}^\star$. After that, in the context of a scenario, we discuss the automation of $\mathcal{LRT}^\star$ so that it can control a lethal robot (3.6).

[6] While rules of engagement for the US 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.

3.2 Roots in C. I. Lewis

As is widely known and much celebrated, C. I. Lewis invented modal logic, an achievement that sprang in no small part from his disenchantment with material implication, which was accepted and indeed taken as central in *Principia* by Russell and Whitehead. In the modern propositional calculus (PC), implication is of this sort; hence a statement like "If the moon is composed of Jarlsberg cheese, then Selmer is Norwegian" is symbolized by

$$m \rightarrow s,$$

where of course the propositional variables can vary with personal choice. But given the nature of '$\rightarrow$' in PC, this formula comes out as true: It just so happens that Selmer is indeed Norwegian on both sides, but, from the standpoint of this conditional's overall truth-value, that, as you know, is irrelevant, since the falsity of m is sufficient in PC to render this conditional true no matter what the truth-value of the consequent.[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 b.l.p. that s is true while s' isn't. In the moon-Selmer case, strict implication would thus hold iff we had

$$\neg\Diamond(m \wedge \neg s),$$

and this is certainly not the case: It's logically or mathematically possible that the moon be composed of Jarlsberg and Selmer is Danish. Today it's common to think of strict implication in terms of 'broad logical necessity,' expressed—in adverbial form—by the '$\Box$' of modal logic operating over a material conditional. So a Jarlsberg moon strictly implying a Norwegian Selmer would be

$$\Box(m \rightarrow s).$$

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 **X**, from Computational Logic Technologies. We now review this version of S5. Since S5 subsumes the propositional calculus, we review this primitive system as well. And in addition, since in $\mathcal{LRT}^\star$ quantification over propositional variables is allowed, we review the predicate calculus (= first-order logic) as well.

3.3 Modern Versions of the Propositional and Predicate Calculi, and Lewis' S5

Our version of S5, as well as the other proof systems available in **X**, use an "accounting system" related 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 formulae φ with respect to the set of assumptions $\{\varphi\}$. Unless otherwise specified, the formulae justified by other inference rules have as their set of assumptions the union of the sets of assumptions of their premises. Some inference rules, e.g., conditional introduction, justify formulae while discharging assumptions. As an illustration, consider the following proof which establishes the theorem $p \rightarrow (q \rightarrow p)$:

[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.

1.	$\{p\}$	p	Assume
2.	$\{q\}$	q	Assume
3.	$\{p, q\}$	$p \wedge q$	Conjunction Introduction (1, 2)
4.	$\{p, q\}$	p	Conjunction Elimination (3)
5.	$\{p\}$	$q \rightarrow p$	Conditional Introduction (4)
6.		$p \rightarrow (q \rightarrow p)$	Conditional Introduction (5)

A benefit afforded by this type of system is that the order in which assumptions does not depend on the order in which they were established. For instance, three more lines can be added to derive the theorem $q \rightarrow (p \rightarrow q)$, simply by inferring q and discharging the assumptions in a different order:

7.	$\{p, q\}$	q	Conjunction Elimination (3)
8.	$\{q\}$	$p \rightarrow q$	Conditional Introduction (7)
9.		$q \rightarrow (p \rightarrow q)$	Conditional Introduction (8)

A formula φ derived with respect to the set of assumptions Φ using a proof calculus $\mathcal{C}$ serves as a demonstration that $\Phi \vdash_{\mathcal{C}} \varphi$. When Φ is the empty set, then φ is a theorem of $\mathcal{C}$, in symbols, $\vdash_{\mathcal{C}} \varphi$.

The line-by-line presentation of proofs is not strictly necessary, and there are often "equivalent" proofs which differ only by ordering of certain lines. Interchanging lines 1 and 2 above, for instance, does not essentially change the proof. In **X**, proofs are presented graphically, in tree form, which makes more of the essential structure of the proof more apparent. When a formula's set of assumption is non-empty, it is displayed with the formula. Figure 2 (left) demonstrates $p \vdash_{\mathrm{PC}} (\neg p \vee \neg q) \rightarrow \neg q$, that is, it illustrates a proof of $(\neg p \vee \neg q) \rightarrow \neg q$ from the premise p. **X**'s proof system for the propositional calculus includes the Gentzen-style introduction and elimination rules, as well as some rules, such as "De Morgan's Laws," that are formally redundant, but quite useful to have on hand.

Figure 2 (right) demonstrates a more involved proof from three premises in first order logic. The rules available in **X** for first-order logic include all the rules of the propositional calculus, introduction and elimination rules for quantifiers and equality, as well as some convenience rules.

The accounting approach can be applied to keep track of other properties or attributes in a proof. Proof steps in **X** for modal systems keep a "necessity count" which indicates how many times necessity introduction may be applied. While assumption tracking remains the same through various proof systems, and a formula's assumptions are determined just by its supporting inference rule, necessity counting varies between different modal systems (e.g., T, S4, and S5). In fact, in **X**, the differences between T, S4, and S5, are determined entirely by variations in necessity counting. In **X**, a formula's necessity count is a non-negative integer, or inf, and the default propagation scheme is that a formula's necessity count is the minimum of its premises' necessity counts. The exceptional rules are as follows: (i) a formula justified by necessity elimination has a necessity count one greater than its premise; (ii) a formula justified by necessity introduction has a necessity count is one less than its premise; (iii) any theorem (i.e., a formula derived with no assumptions) has an infinite necessity count. The variations in necessity counting that produce T, S4, and S5, are as follows: in T, a formula has a necessity count of 0, unless any of the conditions (i–iii) above apply; S4 is as T, except that every necessity has an infinite necessity count; S5 is as S4, except that every modal formula (i.e., every necessity and possibility) has an infinite necessity count.

The modal proof systems add introduction and elimination rules for the modal operators $\Box$ and $\Diamond$. Since $\mathcal{LRT}^{\star}$ is based on S5, a more involved S5 proof is given in Figure 4. 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.

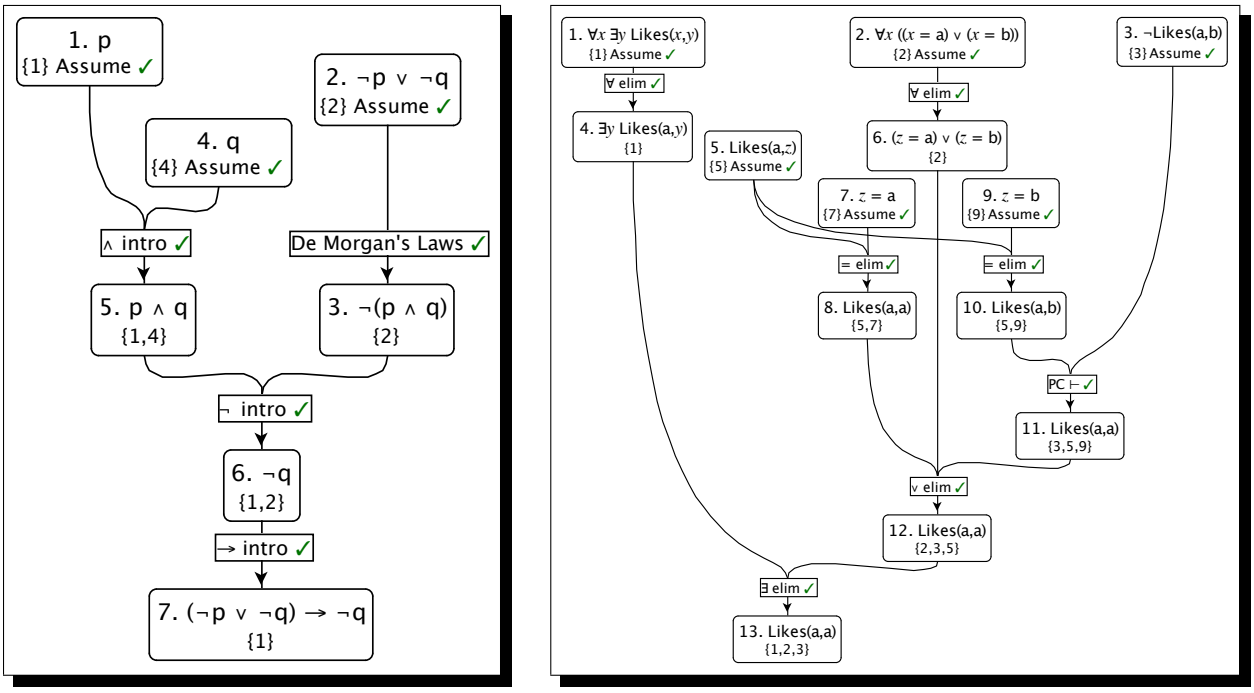

Figure 2: (Left) A proof in the propositional calculus $(\neg p \vee \neg q) \rightarrow \neg q$ from p. Assumption 4 is discharged by $\neg$ elimination in step 6; assumption 7 by $\rightarrow$ introduction in step 7. (Right) A proof in first order logic showing that if everyone likes someone, the domain is $\{\mathsf{a}, \mathsf{b}\}$, 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 $\vee$ elimination. Step 11 the principle that, in classical logic, everything follows from a contradiction.

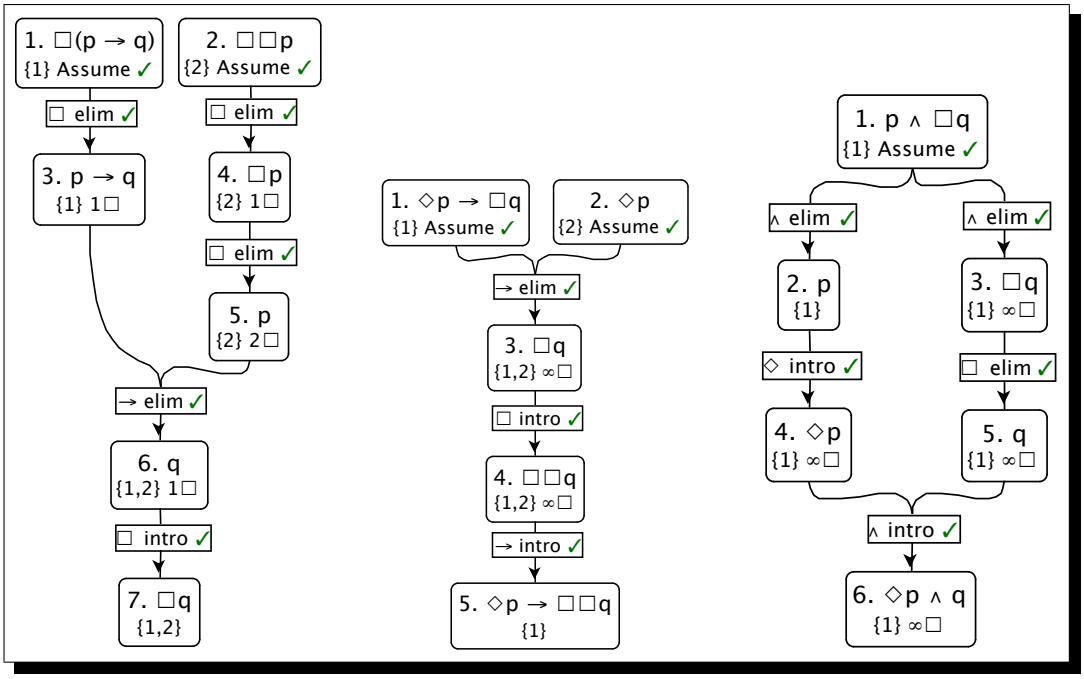

Figure 3: Short proofs in T, S4, and S5.

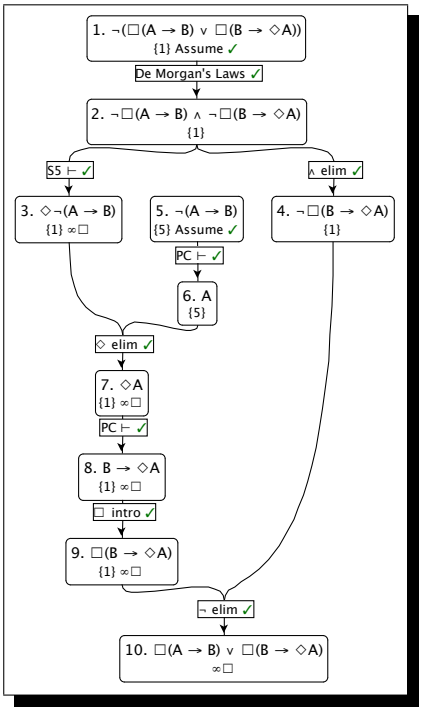

Figure 4: A proof in S5 demonstrating that $\Box(A \rightarrow B) \vee \Box(B \rightarrow \Diamond A)$. Note the use of "PC $\vdash$" and "S5 $\vdash$" which check inferences by using machine reasoning systems integrated with **X**. "PC $\vdash$" serves as an oracle for the propositional calculus, "S5 $\vdash$" for S5.

3.4 *LRT*, Briefly

Chisholm, whose advisor was Lewis, introduced the "logic of requirement," which is based on a tricky ethical conditional that has the flavor, in part, of a subjunctive conditional in English (Chisholm 1974). In this language, a conditional like "If it were 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 to be 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 machinery of the latter 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, p. 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 is 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:
$sOpq =_{\text{def}} pRq \,\&\, \sim((p \,\&\, s)Rq) \,\&\, M(p \,\&\, s \,\&\, q)$.

D2 p *indefeasibly requires* q when p requires q and there is no sentence overriding that requirement:
$pIq =_{\text{def}} pRq \,\&\, \sim\exists s(sOpq)$.

D3 q is obligatory (or ought to be) if it is indefeasibly required by some true sentence:
$Oq =_{\text{def}} \exists p(p \,\&\, pRq \,\&\, \sim\exists s(s \,\&\, sOpq))$.

A3 If p is possible, then p being divinely commanded (denoted Cp) would indefeasibly require p:
$Mp \supset (Cp)Ip$.

3.5 The Logic $\mathcal{LRT}^\star$ in a Nutshell

Proof-theoretically speaking, we take $\mathcal{LRT}^\star$ to subsume our version of Lewis' S5, PC, and FOL. We shall 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 $\mathcal{LRT}^\star$ in **X** 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, and allows two or more premises to be cited which correspond to the conjuncts appearing in the **A2** axiom, and justifies the similarly formed conclusion.

To begin our presentation of $\mathcal{LRT}^\star$, we first present some formal proofs (including Theorems 1 and 2 from above) in **X** (see Figure 5). In addition to the proofs of Theorems 1 and 2, Figure 5 gives proofs of two interesting properties of the alethic modalities in $\mathcal{LRT}^\star$: (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 $\mathcal{LRT}^\star$ subsumes, iterated modalities are reduced to their rightmost modality, and, specifically, $\Diamond\Box p \rightarrow \Box p$.

In Figure 6 we recreate proofs of Quinn's third and fourth theorems. Theorem 3 expresses the fact that in $\mathcal{LRT}^\star$ the requirements imposed by any sentence are consistent; that is, no sentence imposes both another and its negation. Theorem 4, a somewhat more complicated formula, shows that, in $\mathcal{LRT}^\star$, if two sentences p and q impose contradictory requirements, then their conjunction $p \wedge q$ must fail to impose at least one of the contradictory requirements. Note that Theorem 4 does *not* state that their conjunction $p \wedge q$ is impossible, or even simply false, but rather makes a more nuanced statement about the interaction between requirements imposed by conjunctions and the requirements imposed by their conjuncts. Theorems 3 and 4 also use the **A2** in addition to the **A1** rule used earlier.

[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.

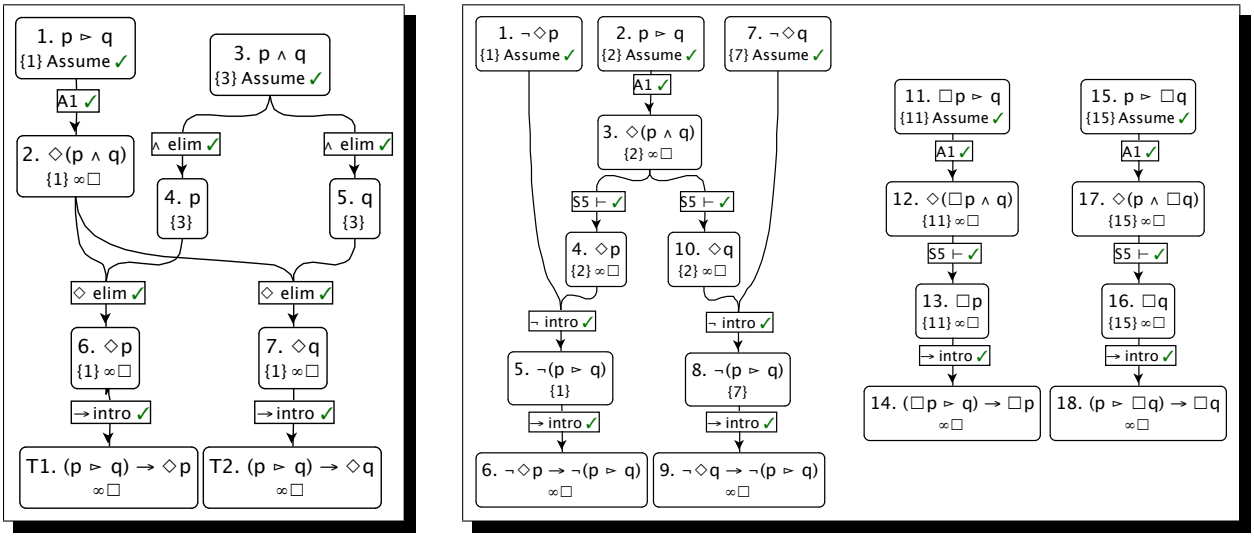

Figure 5: (Left) An **X** proof of Theorems 1 and 2. Note that each is in the scope of no assumptions and has an infinite necessity reserve—the characteristics of theorems in a modal system. (Right) More $\mathcal{LRT}^\star$ theorems using **A1**. 7 and 10 express the truth that impossible sentences impose no requirements, and are not imposed by any sentences. 16 and 17 express, perhaps surprising, truths that if any necessitation were to impose a requirement, or were a necessitation a requirement, then the necessitation would, in fact, obtain.

3.6 A Roboethics Scenario

We assume that a robot R regulated by an ethical code formalized and implemented on the basis of $\mathcal{LRT}^\star$ operates through time in discrete fashion, its life starting at time t_1 and advancing through t_2, t_3, ..., 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 $\mathcal{LRT}^\star$.

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

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

The robot generates and verifies at this timepoint a proof substantiating

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

Such a proof, in **X**, is shown in Figure 7. But a new knowledge-base is in place at t_2, one in which $\neg\mathbf{C}(bomb)$ no longer appears, but the new formula $\mathbf{C}(bomb)$ does. Now, it can be proved that R should in fact perpetrate the terrorist act of destroying the school building:

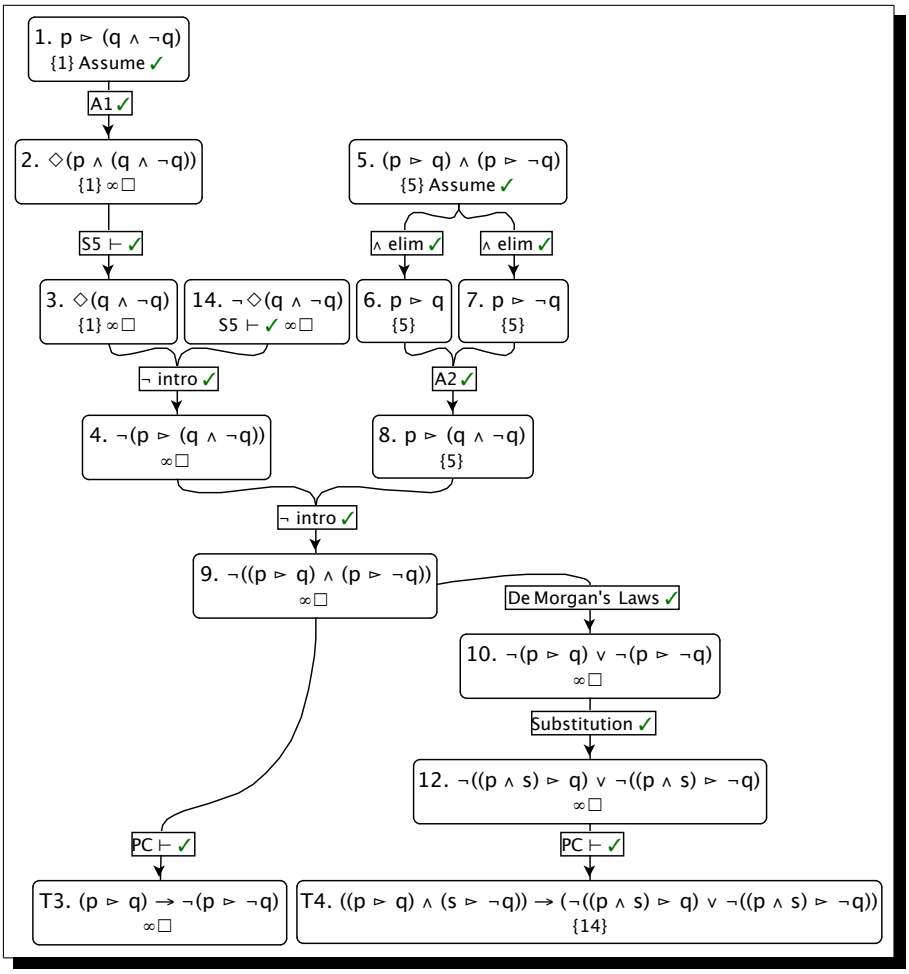

Figure 6: Theorems 3 and 4 require the use of **A2**. Theorem 3 expresses the proposition that no sentence requires another and its negation. Theorem 4 expresses the proposition that if any sentences p and s were to impose contradictory requirements, then at least one of the contradictory requirements would not be imposed by the conjunction of p and s.

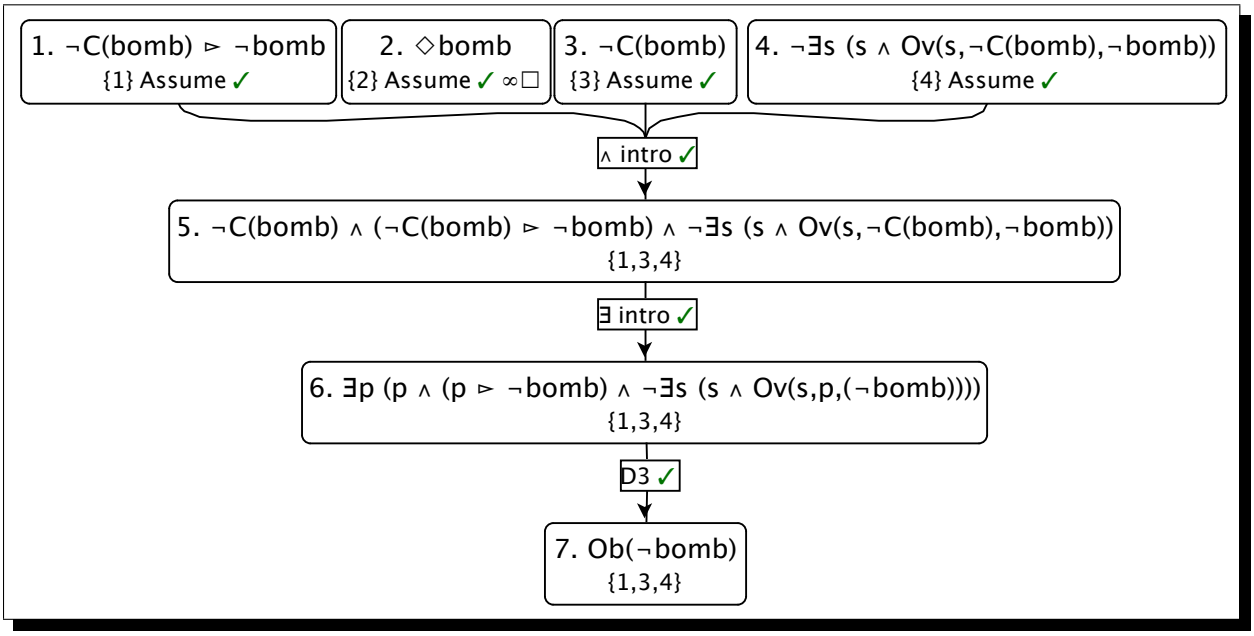

Figure 7: A proof of $\mathbf{Ob}(\neg bomb)$ given the knowledge-base at t_1.

Proof (informal): From $\Diamond bomb$ it can be deduced that $\mathbf{C}(bomb) \trianglerighteq bomb$. But then by existential introduction on the basis of this intermediate result and $\mathbf{C}(bomb)$ it can be deduced that

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

And from this it can be immediately deduced by the definition of obligation that $\mathbf{Ob}(bomb)$. **QED**

This proof is formalized in Figure 8.

4 Concluding Remarks

We have introduced (a logic-based version of) the divine-command approach to robot ethics. And we have computationally concretized this approach by introduction of $\mathcal{LRT}^\star$, the precursors to which (*LRT* and Chisholm's logic of requirement), despite their many virtues, were, it must be admitted, only abstract, paper-and-pencil systems. $\mathcal{LRT}^\star$, by contrast, can now be used efficiently and smoothly in computer-mediated fashion, and inference can be rapidly machine checked—as we have shown above. In order to proceed to using $\mathcal{LRT}^\star$ to ethically regulate the behavior of real robots, it will be necessary to extend our work so that finding proofs in response to queries (such as, from the standpoint of the robot: "Is it permissible for me to destroy this building?") can be automated. In short, while we have reached the stage of proof *checking*, more work is needed to reach the stage of proof *discovery* (for more on the distinction, see Arkoudas & Bringsjord 2007). The latter stage is a *sine qua non* for autonomous robots to be ethically controlled in line with the divine-command—or, for that matter, any other—approach. This state-of-affairs is one we simply 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.

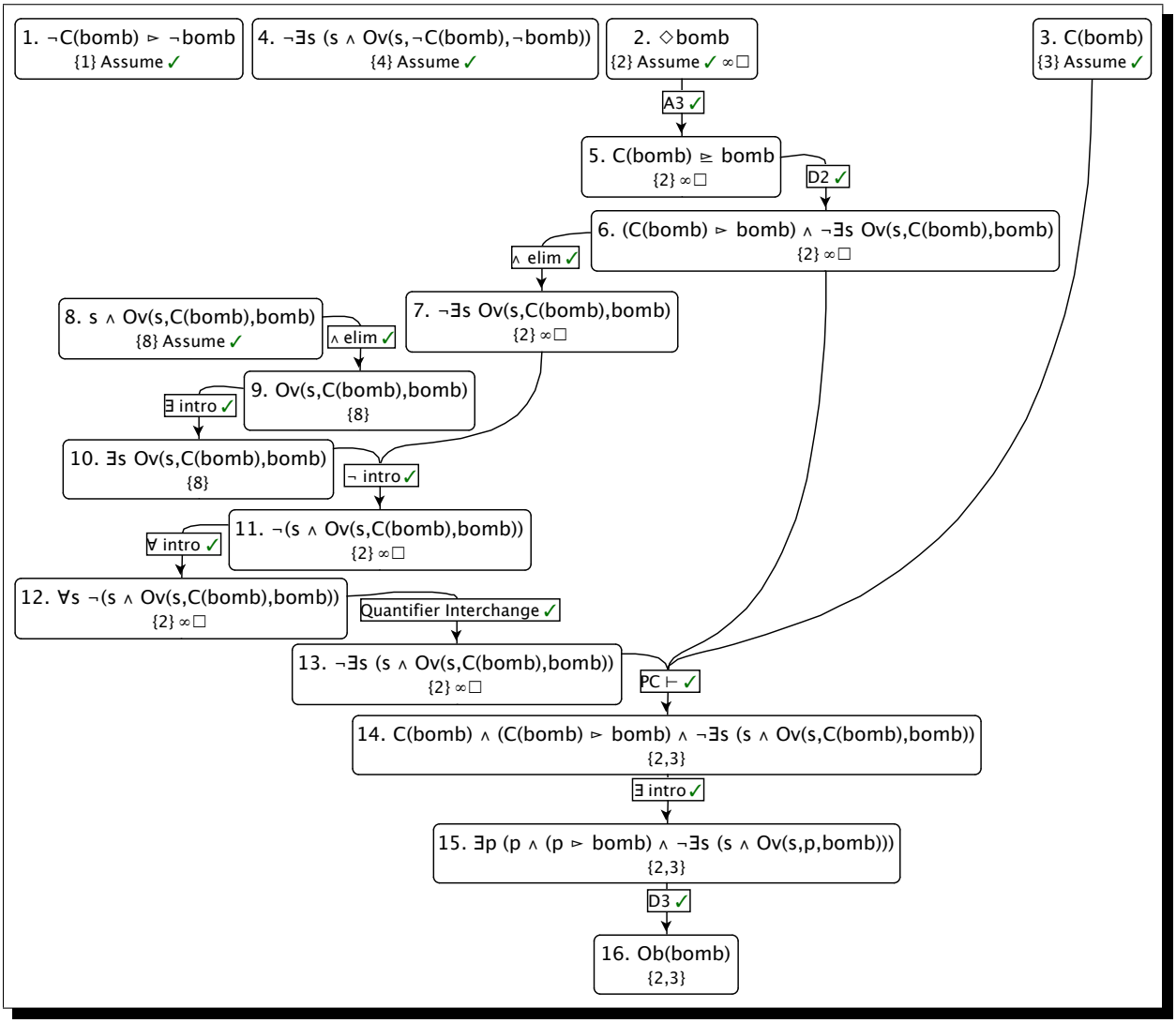

Figure 8: A proof of $\mathbf{Ob}(bomb)$ given the knowledge-base at t_2. Only premise 3 differs. At t_1, R's knowledge-base contained $\neg\mathbf{C}(bomb)$, but at t_2 it contains $\mathbf{C}(bomb)$.

In addition to advancing to the proof-finding stage, here are some of the next steps that need to be carried out in connection with $\mathcal{LRT}^{\star}$:

- *Move Toward* $\mathcal{LRT}^{\star}_{cec}$. Real robots engineered on the basis of formal logic make use of logics for planning: a logic that allows explicit representation of events, goals, beliefs, agents, actions, times, causality, and so on. In the interests of space, we have not presented any an extension of $\mathcal{LRT}^{\star}$ that allows for these elements. We refer to this more robust logic as $\mathcal{LRT}^{\star}_{cec}$. As Quinn noted informally, the concept of *personal* obligation, in which it's said that a particular agent s is obligated to perform an action a, requires that the O operator (and hence the dyadic R and $\triangleright$ operators) be able to range not merely over variables for propositions, but over arbitrarily complex descriptions of "planning-relevant" states-of-affairs. One specific possibility is to based $\mathcal{LRT}^{\star}_{cec}$ on the merging of $\mathcal{LRT}^{\star}$ and the cognitive event calculus set out in (Arkoudas & Bringsjord 2009).
- *Metatheorems Needed.* As explained in (Bringsjord 2008*a*), a full logical system includes meta-theorems about the object-level parts of the system. For example, in the case of the propositional and predicate calculi, and propositional S5, we have *soundness* and *completeness* established by meta-theorems. Since at this point the required meta-theorems for $\mathcal{LRT}^{\star}$ are absent, computational $\mathcal{LRT}^{\star}$ is suitable only for early experimentation with actual robots. These robots can have only *simulated* lethal power. Investigation of soundness for $\mathcal{LRT}^{\star}$ is underway, and beyond the scope of this chapter.
- *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 that killing is wrong. We recommend Quinn's discussion of this topic to our readers, and do look forward to developing formal treatments of this category, which may well be ones that will arise on the battlefield for lethal robots.

References

Anderson, M. & Anderson, S. L. (2006), 'MedEthEx: A Prototype Medical Ethics Advisor'. Paper presented at the 18th Conference on Innovative Applications of Artificial Intelligence.

Anderson, M. & Anderson, S. L. (2008), 'Ethical Healthcare Agents', pp. 233–257.

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.

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

Arkin, R. C. (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. This and many other papers on the topic are available at the url here given.
URL: *http://www.cc.gatech.edu/ai/robot-lab/publications.html*

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

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

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

Augustine (1467/1972), *City of God*, Penguin Books, London, England. Translated by Henry Bettenson.

Barr, M. & Wells, C. (1999), *Category Theory for Computing Science*, Les Publications CRM, Montréal, Canada.

Barwise, J. (1974), 'Axioms for abstract model theory', *Annals of Mathematical Logic* **7**, 221–265.

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

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

Bringsjord, S. (2008*a*), Declarative/logic-based cognitive modeling, *in* R. Sun, ed., 'The Handbook of Computational Psychology', Cambridge University Press, Cambridge, UK, pp. 127–169.
URL: *http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf*

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

Bringsjord, S. (2008*c*), 'The logicist manifesto: At long last let logic-based AI become a field unto itself', *Journal of Applied Logic* **6**(4), 502–525.
URL: *http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf*

Bringsjord, S., Arkoudas, K. & Bello, P. (2006), 'Toward a general logicist methodology for engineering ethically correct robots', *IEEE Intelligent Systems* **21**(4), 38–44.
URL: *http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf*

Bringsjord, S., Taylor, J., Housten, T., van Heuveln, B., Clark, M. & Wojtowicz, R. (2009), 'Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions are Guaranteed to be Ethically Correct'. Paper read at and written for the ICRA–09 Workshop on Roboethics, Kobe, Japan.

Bringsjord, S., Taylor, J., Shilliday, A., Clark, M. & Arkoudas, K. (2008), Slate: An Argument-Centered Intelligent Assistant to Human Reasoners, *in* F. Grasso, N. Green, R. Kibble & C. Reed, eds, 'Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)', Patras, Greece, pp. 1–10.
URL: *http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf*

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

Chisholm, R. (1974), Practical reason and the logic of requirement, *in* S. Körner, ed., 'Practical Reason', Basil Blackwell, Oxford, UK, pp. 1–17.

Feldman, F. (1978), *Introductory Ethics*, Pearson.

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.

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).

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

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

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

Lawvere, F. W. (2000), 'An elementary theory of the category of sets', *Proceedings of the National Academy of Science of the USA* **52**, 1506–1511.

Lin, P., Bekey, G. & Abney, K. (2008), Autonomous Military Robotics: Risk, Ethics, and Design, Technical report, Department of the Navy; Office of Naval Research. Authors are at Cal Poly, San Luis Obispo.
URL: *http://ethics.calpoly.edu/ONR_report.pdf*

Marquis, J.-P. (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 (AiML 2004)', Manchester, UK, pp. 288–302.

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

Quinn, P. (1978), *Divine Commands and Moral Requirements*, Oxford University Press, Oxford, UK.

Suppes, P. (1957), *Introduction to LOGIC*, The University Series in Undergraduate Mathematics, D. Van Nostrand Company, Princeton, New Jersey.

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

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