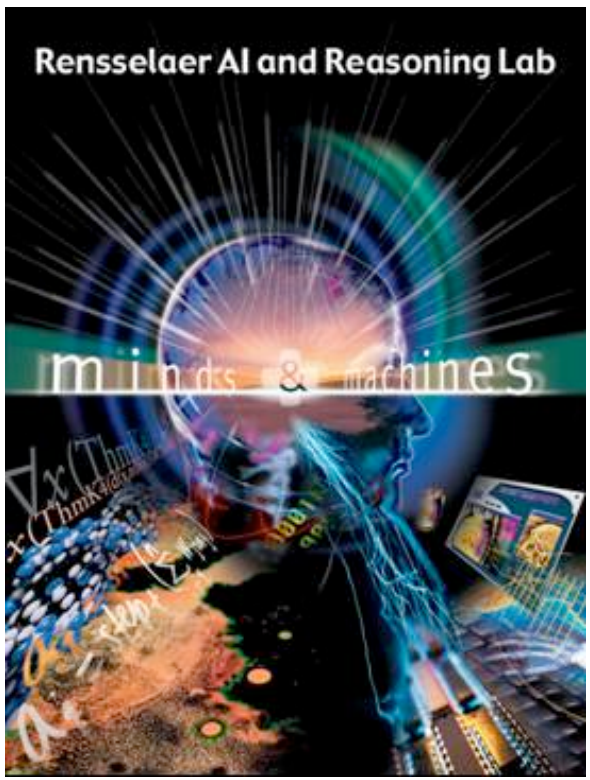
Unethical but Rule-Bound Robots Would Kill Us All



Selmer Bringsjord
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
for *Ethical Robots* @ IU
Future of AI @ AGI 2009 3.9.09

Rensselaer | DEPARTMENT OF COGNITIVE SCIENCE

Rensselaer — Computer Science

Abstract

A robot can flawlessly obey a "moral" code of conduct and still be thoroughly, stupidly, catastrophically unethical. This is easy to prove: Imagine a code of conduct that recommends some action which is, in the broader context, positively immoral. For example, if human Jones has a device which, if not eliminated, will (by his plan) see to the incineration of an metropolis, and a robot (an unmanned, autonomous UAV, e.g.) is bound by a code of conduct not to destroy Jones because he happens to be a civilian, or be in a church, or at a cemetery ... and the robot has just one shot, and this is it, it would be immoral not to eliminate Jones. But unfortunately, the US government is apparently sponsoring work designed to bind robots by codes of conduct (e.g., rules of engagement covering warfighters). This approach is going to get us all killed, as sure as I'm Norwegian. The approach that *won't* get us killed, and indeed the only viable path open to us if we want to survive, is to control robot behavior by fundamental ethical principles expressed in deontic logic and the like — principles from which *suitable* codes can be mechanically *derived* by robots *on the fly*.

Abstract

Abstract-2

A robot can flawlessly obey a "moral" code of conduct and still be thoroughly, stupidly, catastrophically unethical. This is easy to prove: Imagine a code of conduct that recommends some action which is, in the broader context, positively immoral. For example, if human Jones has a device which, if not eliminated, will (by his plan) see to the incineration of an metropolis, and a robot (an unmanned, autonomous UAV, e.g.) is bound by a code of conduct not to destroy Jones because he happens to be a civilian, or be in a church, or at a cemetery ... and the robot has just one shot, and this is it, it would be immoral not to eliminate Jones. But unfortunately, the US government is apparently sponsoring work designed to bind robots by codes of conduct (e.g., rules of engagement covering warfighters). This approach is going to get us all killed, as sure as I'm Norwegian. The approach that *won't* get us killed, and indeed the only viable path open to us if we want to survive, is to control robot behavior by fundamental ethical principles expressed in meta- and integrative logical systems that range over logical systems in which different ethical codes have been represented — systems from which *suitable* codes can be mechanically *derived* by robots *on the fly*.

Abstract-2

Abstract-3

Abstract-3

A robot can flawlessly obey a "moral" code of conduct and still be thoroughly, stupidly, catastrophically unethical. This is easy to prove: Imagine a code of conduct that recommends some action which is, in the broader context, positively immoral. For example, if human Jones has a device which, if not eliminated, will (by his plan) see to the incineration of an metropolis, and a robot (an unmanned, autonomous UAV, e.g.) is bound by a code of conduct not to destroy Jones because he happens to be a civilian, or be in a church, or at a cemetery ... and the robot has just one shot, and this is it, it would be immoral not to eliminate Jones. But unfortunately, the US government is apparently sponsoring work designed to bind robots by codes of conduct (e.g., rules of engagement covering warfighters). This approach is going to get us all killed, as sure as I'm Norwegian. The approach that *won't* get us killed, and indeed the only viable path open to us if we want to survive, is to (1) control robot behavior by fundamental ethical principles expressed in meta- and integrative logical systems that range over logical systems in which different ethical codes have been represented, and (2) (program) verify these systems in unprecedented ways that produce an unprecedentedly high level of confidence in the operation of these systems — and from these systems *suitable* codes can be mechanically *derived* by robots *on the fly*.

The Problem (barbarically put) ...

Our Future

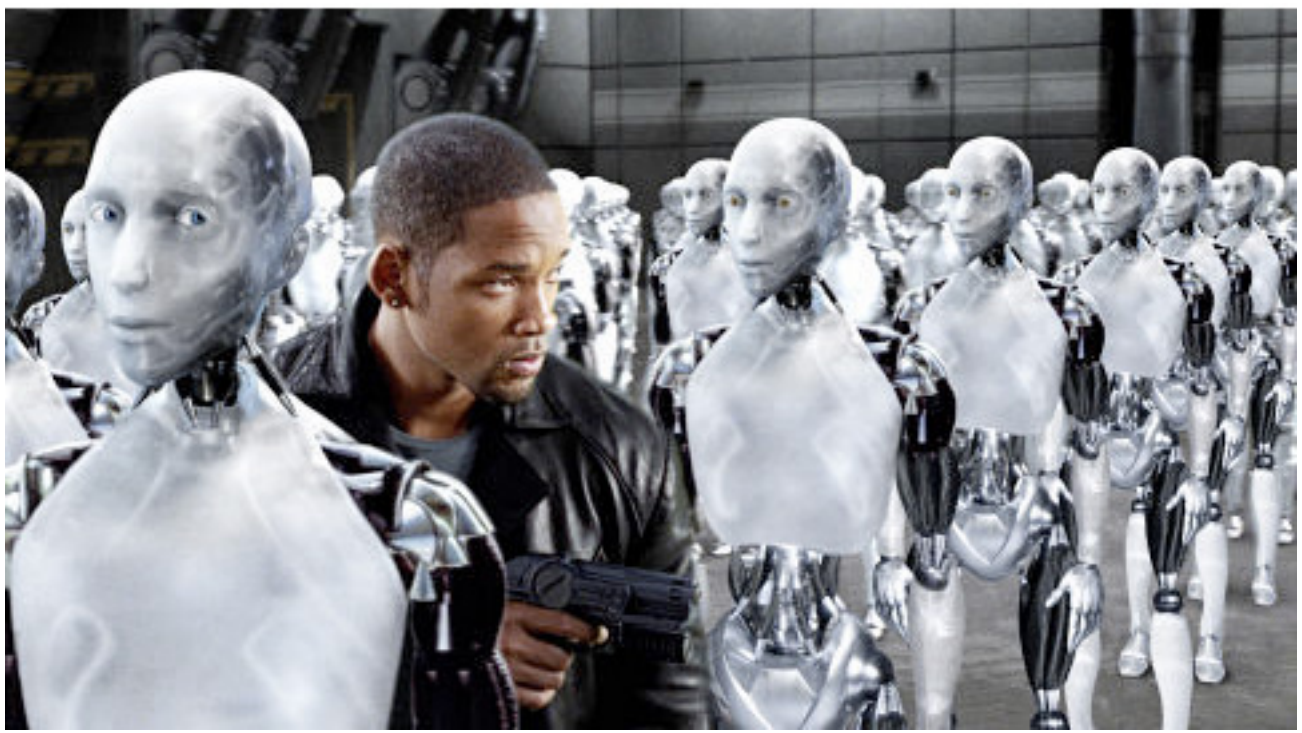
Robots on the battlefield.
Robots in our hospitals.
Robots in law enforcement.
...

Our Problem

If these robots behave immorally, we are killed, or worse.

Our Problem

If these robots behave immorally, we are killed, or worse.



Problem, More Specifically

Problem, More Specifically

- How can we ensure that the robots in question always behave in an ethically correct manner?
- How can we know *ahead of time*, via rationales expressed in clear English (and/or other natural languages), that they will so behave?
- How can we know in advance that their behavior will be constrained specifically by the ethical codes affirmed by human overseers?

Bill Joy:

"We can't."

Bill Joy:

"We can't."

(Bringsjord, S. (2008) "The Future Can Heed Us" *AI & Society*.)

The Solution?

Regulate the behavior of robots with a computational logic, so that all actions they perform are provably ethically permissible.

Toward a General Logicist Methodology for Engineering Ethically Correct Robots

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello,
Rensselaer Polytechnic Institute

A deontic logic formalizes a moral code, allowing ethicists to render theories and dilemmas in declarative form for analysis. It offers a way for human overseers to constrain robot behavior in ethically sensitive environments.

As intelligent machines assume an increasingly prominent role in our lives, there seems little doubt they will eventually be called on to make important, ethically charged decisions. For example, we expect hospitals to deploy robots that can administer medications, carry out tests, perform surgery, and so on, supported by software agents, or softbots, that will manage related data. (Our discussion of ethical robots extends to all artificial agents, embodied or not.) Consider also that robots are already finding their way to the battlefield, where many of their potential actions could inflict harm that is ethically impermissible.

How can we ensure that such robots will always behave in an ethically correct manner? How can we know ahead of time, via rationales expressed in clear natural languages, that their behavior will be constrained specifically by the ethical codes affirmed by human overseers? Pessimists have claimed that the answer to these questions is: "We can't!" For example, Sun Microsystems' cofounder and former chief scientist, Bill Joy, published a highly influential argument for this answer.[1] Inevitably, according to the pessimists, AI will produce robots that have tremendous power and behave immorally. These predictions certainly have some traction, particularly among a public that pays good money to see such dark films as Stanley Kubrick's *2001* and his joint venture with Stephen Spielberg, *AI*).

Nonetheless, we're optimists: we think formal logic offers a way to preclude doomsday scenarios of malicious robots taking over the world. Faced with the challenge of engineering ethically correct robots, we propose a logic-based approach (see the related sidebar). We've successfully implemented and demonstrated this approach.[2] We present it here in a general methodology to answer the ethical questions that arise in entrusting robots with more and more of our welfare.

Deontic logics: Formalizing ethical codes

Our answer to the questions of how to ensure ethically correct robot behavior is, in brief, to insist that robots only perform actions that can be proved ethically permissible in a human-selected *deontic logic*. A deontic logic formalizes an ethical code—that is, a collection of ethical rules and principles. Isaac Asimov introduced a simple (but subtle) ethical code in his famous Three Laws of Robotics:[3]

1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Human beings often view ethical theories, principles, and codes informally, but intelligent machines require a greater degree of precision. At present, and for the foreseeable future, machines can't work directly with natural language, so we can't simply feed Asimov's three laws to a robot and instruct it behave in

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

Toward a General Logicist Methodology for Engineering Ethically Correct Robots

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello,
Rensselaer Polytechnic Institute

A deontic logic formalizes a moral code, allowing ethicists to render theories and dilemmas in declarative form for analysis. It offers a way for human overseers to constrain robot behavior in ethically sensitive environments.

As intelligent machines assume an increasingly prominent role in our lives, there seems little doubt they will eventually be called on to make important, ethically charged decisions. For example, we expect hospitals to deploy robots that can administer medications, carry out tests, perform surgery, and so on, supported by software agents, or softbots, that will manage related data. (Our discussion of ethical robots extends to all artificial agents, embodied or not.) Consider also that robots are already finding their way to the battlefield, where many of their potential actions could inflict harm that is ethically impermissible.

How can we ensure that such robots will always behave in an ethically correct manner? How can we know ahead of time, via rationales expressed in clear natural languages, that their behavior will be constrained specifically by the ethical codes affirmed by human overseers? Pessimists have claimed that the answer to these questions is: "We can't!" For example, Sun Microsystems' cofounder and former chief scientist, Bill Joy, published a highly influential argument for this answer.[1] Inevitably, according to the pessimists, AI will produce robots that have tremendous power and behave immorally. These predictions certainly have some traction, particularly among a public that pays good money to see such dark films as Stanley Kubrick's *2001* and his joint venture with Stephen Spielberg, *AI*).

Nonetheless, we're optimists: we think formal logic offers a way to preclude doomsday scenarios of malicious robots taking over the world. Faced with the challenge of engineering ethically correct robots, we propose a logic-based approach (see the related sidebar). We've successfully implemented and demonstrated this approach.[2] We present it here in a general methodology to answer the ethical questions that arise in entrusting robots with more and more of our welfare.

Deontic logics: Formalizing ethical codes

Our answer to the questions of how to ensure ethically correct robot behavior is, in brief, to insist that robots only perform actions that can be proved ethically permissible in a human-selected *deontic logic*. A deontic logic formalizes an ethical code—that is, a collection of ethical rules and principles. Isaac Asimov introduced a simple (but subtle) ethical code in his famous Three Laws of Robotics:[3]

1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Human beings often view ethical theories, principles, and codes informally, but intelligent machines require a greater degree of precision. At present, and for the foreseeable future, machines can't work directly with natural language, so we can't simply feed Asimov's three laws to a robot and instruct it behave in

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

The Solution?

Solution Steps

Solution Steps

1. Human overseers select ethical theory, principles, rules.

Solution Steps

1. Human overseers select ethical theory, principles, rules.
2. Selection is formalized in a deontic logic, revolving around what is permissible, forbidden, obligatory (etc).

Solution Steps

1. Human overseers select ethical theory, principles, rules.
2. Selection is formalized in a deontic logic, revolving around what is permissible, forbidden, obligatory (etc).
3. The deontic logic is mechanized.

Solution Steps

1. Human overseers select ethical theory, principles, rules.
2. Selection is formalized in a deontic logic, revolving around what is permissible, forbidden, obligatory (etc).
3. The deontic logic is mechanized.
4. Every action that is to be performed must be provably ethically permissible relative to this mechanization (with all proofs expressible in smooth English).

Simple Example...

Context

- The year is 2020.
- Health care is delivered in large part by interoperating teams of robots and softbots.
- Hospital ICU.
- Robot R_1 caring for H_1; R_2 for H_2.
- H_1 on life support.
- H_2 stable, but in desperate need of expensive pan med.

More Context

- Two actions performable by the robotic duo of R1 and R2, both of which are rather unsavory, ethically speaking:
  - *term*
  - *delay*

Encapsulation

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

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

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

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

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

$$\vdots$$

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

where $C = O^\star$

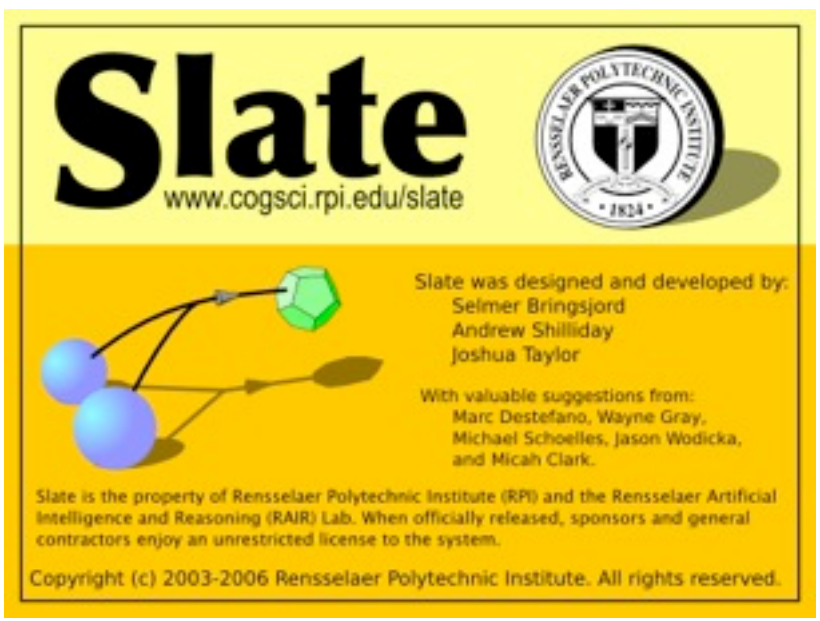
But There is a Twist

But There is a Twist

- It is: An *interactive* reasoning system is required.
  - Examples of such systems include Athena, and Slate.
- Human consultation and assistance must be provided, because machines are such dim reasoners.

But There is a Twist

- It is: An *interactive* reasoning system is required.
  - Examples of such systems include Athena, and Slate.
- Human consultation and assistance must be provided, because machines are such dim reasoners.



This won't work. We will be killed.

Program Verification ...

Minds & Machines (2007) 17:185–202
DOI 10.1007/s11023-007-9063-5

Computers, Justification, and Mathematical Knowledge

Konstantine Arkoudas · Selmer Bringsjord

Published online: 23 June 2007

Abstract The original proof of the four-color theorem by Appel and Haken sparked a controversy when Tymoczko used it to argue that the justification provided by unsurveyable proofs carried out by computers cannot be a priori. It also created a lingering impression to the effect that such proofs depend heavily for their soundness on large amounts of computation-intensive custom-built software. Contra Tymoczko, we argue that the justification provided by certain computerized mathematical proofs is not fundamentally different from that provided by surveyable proofs, and can be sensibly regarded as a priori. We also show that the aforementioned impression is mistaken because it fails to distinguish between proof search (the context of discovery) and proof checking (the context of justification). By using mechanized proof assistants capable of producing certificates that can be independently checked, it is possible to carry out complex proofs without the need to trust arbitrary custom-written code. We only need to trust one fixed, small, and simple piece of software: the proof checker. This is not only possible in principle, but is in fact becoming a viable methodology for performing complicated mathematical reasoning. This is evinced by a new proof of the four-color theorem that appeared in 2005, and which was developed and checked in its entirety by a mechanical proof system.

Keywords A priori · Justification · Proofs · Certificates · Four-color theorem · Mathematical knowledge

K. Arkoudas (✉) · S. Bringsjord
Cognitive Science Department, Computer Science Department, RPI, Troy, NY, USA 12180
e-mail: arkouk@rpi.edu

S. Bringsjord
e-mail: brings@rpi.edu

Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions are Guaranteed to be Ethically Correct

Selmer Bringsjord, Joshua Taylor
Trevor Housten, Bram van Heuveln
Konstantine Arkoudas, Micah Clark
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Ralph Wojtowicz
Metron Inc.
1818 Library Street
Suite 600
Reston VA 20190 USA

I. Introduction

This is an extended abstract, not a polished paper; an *approach* to, rather than the results of, sustained research and development in the area of roboethics is described herein. Encapsulated, the approach is to engineer ethically correct robots by giving them the capacity to reason *over*, rather than merely *in*, logical systems (where logical systems are used to formalize such things as ethical codes of conduct for warfighting robots). This is to be accomplished by taking seriously Piaget's position that sophisticated human thinking exceeds even abstract processes carried out *in* a logical system, and by exploiting category theory to render in rigorous form, suitable for mechanization, structure-preserving mappings that Bringsjord, an avowed Piagetian, sees to be central in rigorous and rational human ethical decision-making.

We assume our readers to be at least somewhat familiar with elementary classical logic and category theory. Introductory coverage of the former subject can be found in [1], [2];[1] such coverage of the latter, offered from a suitably computational perspective, is provided in [3]. Additional references are of course provided in the course of this document.

II. Piaget's View of Thinking

Many people, including many outside psychology and cognitive science, know that Piaget seminally — and by Bringsjord's lights, correctly — articulated and defended the view that mature human reasoning and decision-making consists in processes operating for the most part on formulas in the language of classical extensional logic (e.g., see [4]).[2] You may yourself have this knowledge. You may also know that Piaget posited a sequence of cognitive stages through which humans, to varying degrees, pass. How many stages are there, according to Piaget? The received answer is: four; and in the fourth and final one, *formal operations*, neurobiologically normal humans can reason accurately and quickly over formulas expressed in the logical system known as first-order logic ($\mathcal{L}_I$).[3]

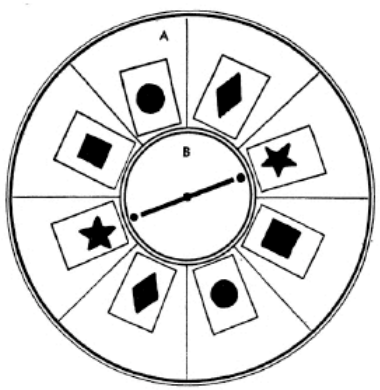


Fig. 1. Piaget's famous "rigged" rotating board to test for the development of Stage-3-or-better reasoning in children. The board, A, is divided into sectors of different colors and equal surfaces; opposite sectors match in color. B is a rotating disk with a metal rod spanning its diameter — but the catch is that the star cards have magnets buried under them (inside wax), so the alignment after spinning is invariably as shown here, no matter how the shapes are repositioned in the sectors (with matching shapes directly across from each other). This phenomenon is what subjects struggle to explain. Details can be found in [4].

Judging by the cognition taken by Piaget to be stage-three or stage-four (e.g., see Figure 1, which shows one

[1] Online, elegant, economical coverage can be found at http://plato.stanford.edu/entries/logic-classical/

[2] Many readers will know that Piaget's position long ago came under direct attack, by such thinkers as Wason and Johnson-Laird [5], [6]. In fact, unfortunately, for the most part people believe that this attack succeeded. Bringsjord doesn't agree in the least, but this isn't the place to visit the debate in question. Interested readers can consult [7], [8].

[3] Various other symbols are used, e.g., the more informative $\mathcal{L}_{\omega\omega}$.

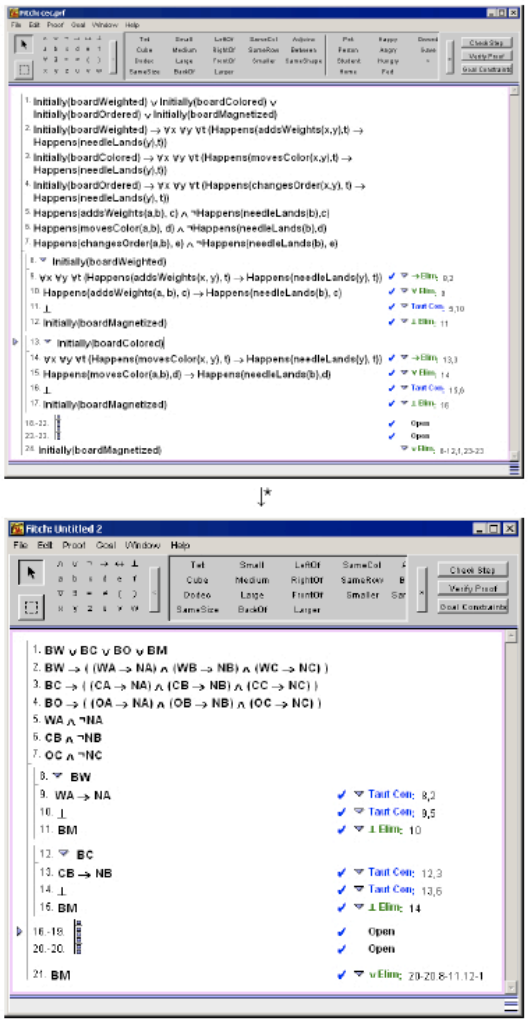


Fig. 2. This figure shows two proofs, one expressed in $\mathcal{L}_I$, the other in $\mathcal{L}_{PC}$. The first-order proof produces the conclusion that what causes the metal rod to invariably stop at the stars is that there are hidden magnets. The basic structure of this proof is proof by cases. Of the four disjuncts entertained as the possible source of the rod-star regularity, the right one is deduced when the others are eliminated. The functor $\star$ is shown here to indicate that the basic structure can be preserved in a proof couched exclusively in the propositional calculus.

see when deployed in warfare and counter-terrorism, where post-stage-four reasoning and decision-making is necessary for successfully handling these situations. The work here is connected to NSF-sponsored efforts on our part to extend CMU's Tekkotsu [20], [21] framework so that it includes operators that are central to our logicist approach to robotics, and specifically to roboethics — for example, operators for belief (**B**), knowledge (**K**), and obligation (◯ of standard deontic logic). The idea is that these operators would link to their counterparts in bona fide calculi for automated and semi-automated machine reasoning. One such calculus has already been designed and implemented: the *socio-cognitive calculus*; see [22]. This calculus includes the full event calculus.

Given that our initial experiments will make use of simple hand-eye robots recently acquired by the RAIR Lab from the Tekkotsu group at CMU, Figure 3, which shows one of these robots, sums up the situation (in connection with the magnet challenge). If sufficiently intricate manipulation cannot be achieved with the simple hand-eye robots, we will use the more powerful PERI, shown in Figure 4.

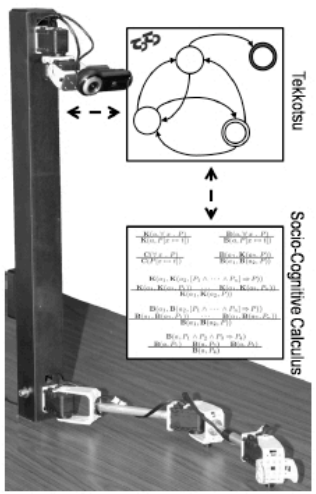
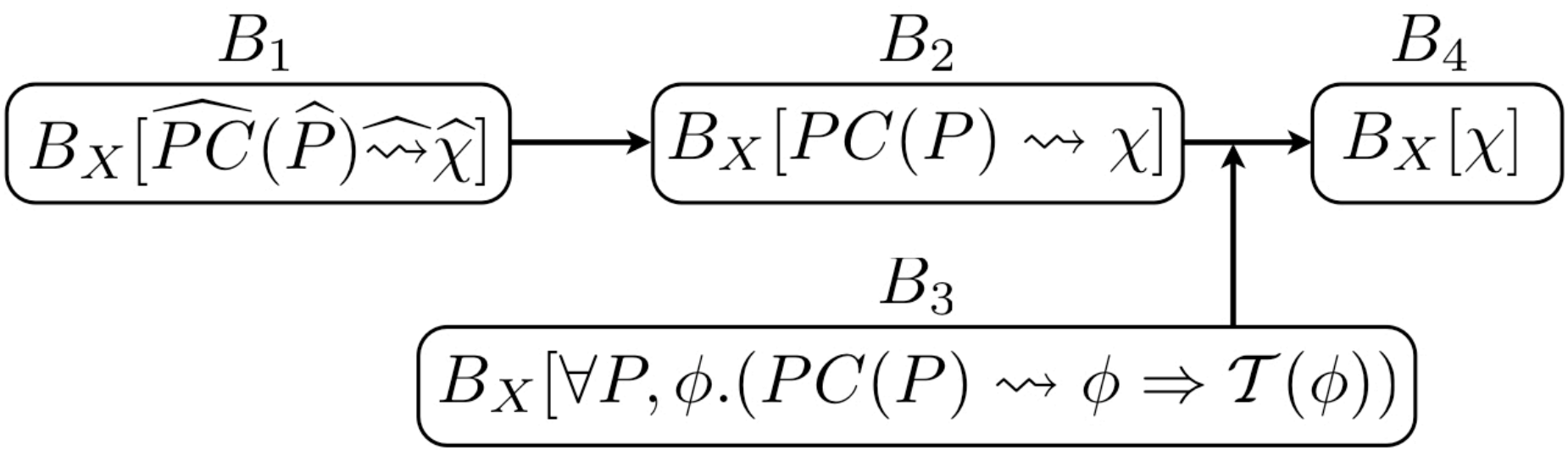


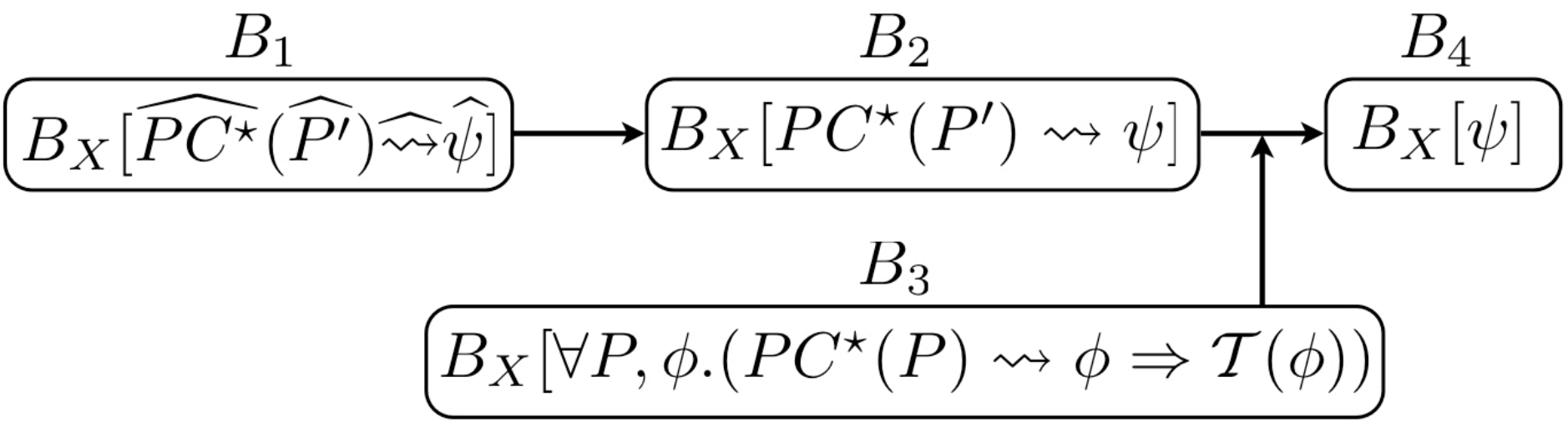
Fig. 3. The basic configuration for our initial implementations.

Fig. 4. The RAIR Lab's PERI

Believing the Completeness of FOL

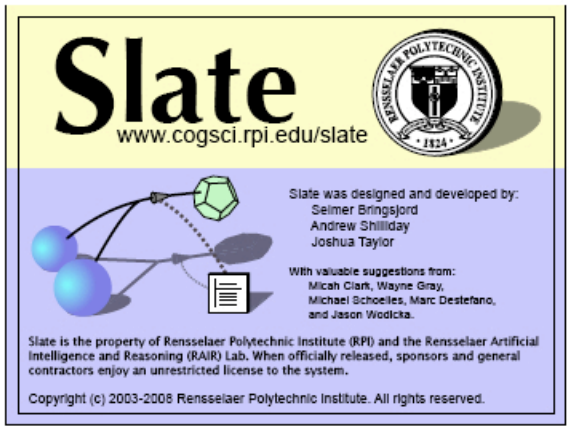


Program Verification



Slate: An Argument-Centered Intelligent Assistant to Human Reasoners

Selmer Bringsjord and Joshua Taylor and Andrew Shilliday
and Micah Clark and Konstantine Arkoudas[1]



Abstract. We describe Slate, a logic-based, robust interactive reasoning system that allows human "pilots" to harness an ensemble of intelligent agents in order to construct, test, and express various sorts of natural argumentation. Slate empowers students and professionals in the business of producing argumentation, e.g., mathematicians, logicians, intelligence analysts, designers and producers of standardized reasoning tests. We demonstrate Slate in several examples, describe some distinctive features of the system (e.g., reading and generating natural language, immunizing human reasoners from "logical illusions"), present Slate's theoretical underpinnings, and note upcoming refinements.

1 INTRODUCTION

Slate is a robust interactive reasoning system. It allows the human "pilot" to harness an ensemble of intelligent agents in order to construct, test, and express natural argumentation of various sorts. Slate is designed to empower students and professionals in the business of producing argumentation, e.g., mathematicians, logicians, intelligence analysts, designers and producers of standardized reasoning tests, and so on. While other ways of pursuing AI may well be preferable in certain contexts, faced with the challenge of having to engineer a system like Slate, a logic-based approach [9, 10, 18, 31, 13] seemed to us ideal, and perhaps the power of Slate even at this point (version 3) confirms the efficacy of this approach. In addition, there is of course a longstanding symbiosis between argumentation and logic revealed in contemporary essays on argumentation [48]. In this paper, we summarize Slate through several examples, describe some distinctive features of the system (e.g., its capacity to read and generate natural language, and to provide human reasoners with apparent immunity from so-called "logical illusions"), say a bit about Slate's theoretical underpinnings, and note upcoming refinements.

[1] Rensselaer Polytechnic Institute (RPI), USA, email: {selmer, tayloj, shilla, clarkm5, arkouk}@rpi.edu

2 A SIMPLE EXAMPLE

We begin by following a fictitious user, Ulric, as he uses Slate to solve a short logic puzzle, the *Dreadsbury Mansion Mystery* [34]:[2]

> Someone who lives in Dreadsbury Mansion killed Aunt Agatha. Agatha, the butler, and Charles live in Dreadsbury Mansion, and are the only people who live therein. A killer always hates his victim, and is never richer than his victim. Charles hates no one that Aunt Agatha hates. Agatha hates everyone except the butler. The butler hates everyone not richer than Aunt Agatha. The butler hates everyone Agatha hates. No one hates everyone. Agatha is not the butler. *Who killed Agatha?*

Information can enter Slate in a number of formats, e.g., as formulae in many-sorted logic (MSL), or as sentences in a logically-controlled English (§4.2). Information can also be imported from external repositories such as databases or the Semantic Web (§4.5). Ulric examines the Dreadsbury Mansion Mystery facts displayed in Slate's workspace (Figure 1).

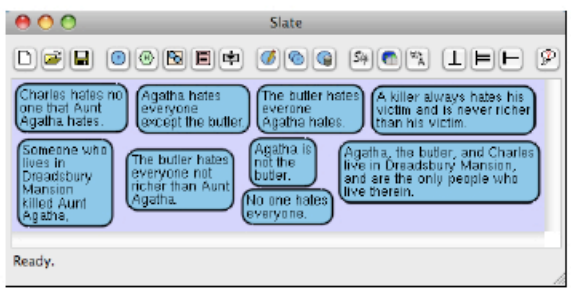


Figure 1. The Dreadsbury Mansion Mystery facts represented in Slate.

A fan of murder mysteries, he considers whether conventional wisdom might hold true, i.e., that the butler did it. Ulric adds the hypothetical to Slate's workspace and asks Slate to check whether the hypothesis is consistent with the other propositions. Slate quickly reports an inconsistency (Figure 2).

[2] This puzzle is of a type typically used to challenge humans (e.g., students in introductory logic courses) and machines (e.g., automated theorem provers).

a Gödelian logic puzzle that approximates GI and demonstrates the power of Slate within demanding logico-mathematical domains like those in which Gödel worked.

A Precursor Gödelian Puzzle. Suppose a machine $\mathcal{M}$ operates on expressions: finite, non-empty sequences of the four glyphs $\sim$, $\star$, P, and M. These four glyphs have intuitive meanings: $\sim$ stands for 'not,' $\star$ for 'to be' or 'is,' P for 'provable,' and M for 'mirror of,' where the mirror of an expression ϕ is the expression $\phi\star\phi$. A sentence is an expression of a particular form, also with an intuitive meaning, specifically,

$P\star\phi$ means that ϕ is provable and is true if and only if ϕ is provable by $\mathcal{M}$.

$PM\star\phi$ means that the mirror of ϕ is provable, and is true if and only if the mirror of ϕ is provable by $\mathcal{M}$.

$\sim P\star\phi$ means that ϕ is not provable, and is true if and only if ϕ is not provable by $\mathcal{M}$.

$\sim PM\star\phi$ means that the mirror of ϕ is not provable, and is true if and only if the mirror of ϕ is not provable by $\mathcal{M}$.

$\mathcal{M}$ is such that it only proves true sentences and never false sentences (i.e., the machine is *sound*). Prove that $\mathcal{M}$ cannot prove all true sentences—there is a true sentence which cannot be proved by $\mathcal{M}$ (i.e., the machine is *incomplete*).

Formalization of the Gödelian Puzzle. We formalize the above puzzle as a logical language consisting of the constants: $\sim$, $\star$, P, M; the (unary) predicates: glyph, expression, sentence, provable, and true; and the functions: cat (concatenation), and mirror. For convenience, we describe as glyphs, expressions, sentences, provable, and true any terms on which glyph, expression, sentence, provable, and true holds, respectively, and denote the application of cat to two terms ϕ and ψ as the concatenation of ϕ and ψ, or by $\phi\psi$, and the application of mirror to a term ϕ as the mirror of ϕ. The interpretation of this vocabulary is subject to the following twelve axioms:

1. *The constants $\sim$, $\star$, P, and M are each distinct.*
2. *The constants $\sim$, $\star$, P, and M are the only glyphs.*
3. *The concatenation of two terms is an expression if and only if both terms are themselves expressions.*
4. *Concatenation is associative.*
5. *The term ϕ is an expression if and only if ϕ is a glyph or is the concatenation of two expressions.*
6. *The mirror of an expression ϕ is defined as the concatenation of ϕ, $\star$, and ϕ (i.e., $\phi\star\phi$).*
7. *If ϕ is an expression, then $P\star\phi$, $PM\star\phi$, $\sim P\star\phi$, and $\sim PM\star\phi$ are sentences.*
8. *If ϕ is an expression then the sentence $P\star\phi$ is true if and only if ϕ is provable.*
9. *If ϕ is an expression, then the sentence $PM\star\phi$ is true if and only if the mirror of ϕ is provable.*
10. *If ϕ is an expression, then the sentence $\sim P\star\phi$ is true if and only if ϕ is not provable.*
11. *If ϕ is an expression, then the sentence $\sim PM\star\phi$ is true if and only if the mirror of ϕ is not provable.*
12. *Every sentence ϕ that is provable is also true.*

The given axioms (propositions 1–12) are represented visually in the Slate workspace in Figure 10, each consisting of the first-order formula derived from the English descriptions above. Moreover, a new intermediate hypothesis is introduced toward the desired goal, viz., that there is a true sentence that cannot be proved by $\mathcal{M}$:

13. *$\sim PM$ is an expression.*

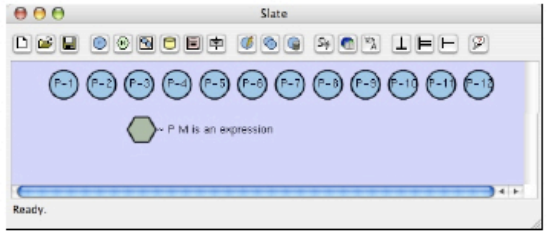


Figure 10. Propositions 1–12 and hypothesis 13 in the Slate workspace.

We indicate that hypothesis 13 is a logical consequence of propositions 2, 3 and 5 by drawing a deductive inference from each of these propositions to hypothesis 13 (Figure 11). Slate is then able to confirm or refute the added inference. Slate does indeed confirm that hypothesis 13 follows from the indicated propositions, by producing as evidence a formal proof which is added to the workspace as a *witness*. Witnesses are objects in Slate that support or weaken inferences. The double-plus symbol indicates that the witness confirms the argument, an ability reserved only for formal proofs. If the inference had been invalid, Slate might have produced a countermodel demonstrating the inference's invalidity.

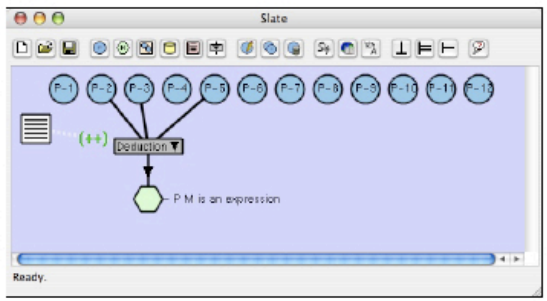


Figure 11. Proof of $\{2,3,5\}\vdash 13$ in the workspace and verified by Slate.

Having proved *$\sim PM$ is an expression*, it follows from 13 and 7 that:

14. *$\sim PM\star\sim PM$ is a sentence.*

If we suppose that *$\sim PM\star\sim PM$ is not true* then by 11 *the mirror of $\sim PM$ is provable* and thus by 6 *$\sim PM\star\sim PM$ is provable*. But then, according to 13 and 14, *$\sim PM\star\sim PM$ is true*—which is in contradiction with our supposition that *$\sim PM\star\sim PM$ is not true*. And so it must be the case that *$\sim PM\star\sim PM$ is true*. In other words, as shown in Figure 12, the hypothesis that

15. *$\sim PM\star\sim PM$ is true.*

follows from axioms 6 and 11 and hypotheses 12 and 13. Since *$\sim PM\star\sim PM$ is true*, it follows from 6 and 11 that

16. *$\sim PM\star\sim PM$ is not provable.*

and consequently, that *there is a true sentence which cannot be proved* (Figure 13).

5.3 Informal Reasoning

When using Slate, the reasoner is able to construct arguments that more closely resemble the uncertain and informal nature of everyday, natural inference. Moreover, the user benefits from the system's

Strength Factors

- Certain
- Evident
- Beyond Reasonable Doubt
- Likely
- Counterbalanced
- ... (symmetrical)

New Question

What could possibly be an alternative approach to solving the problem?

Logic is Our Only Hope

We only have one way to fix the meaning of programs, to verify that they will behave as advertised.

We only have one way to rigorously set out ethical principles.

Enumerative induction will get us killed.

Logic is our only hope, ladies and gentlemen.

Logic is Our Only Hope

We only have one way to fix the meaning of programs, to verify that they will behave as advertised.

We only have one way to rigorously set out ethical principles.

Enumerative induction will get us killed.

Logic is our only hope, ladies and gentlemen.

Finis