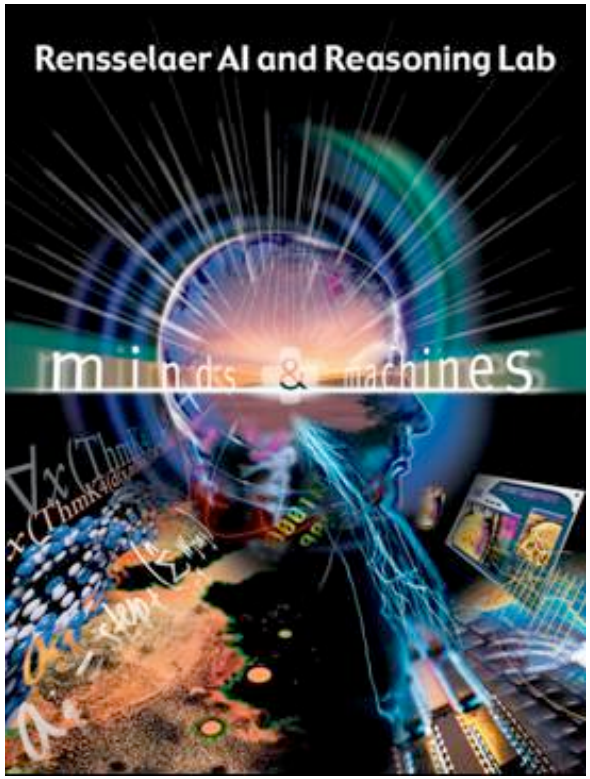
Only a Technology Triad Can Tame Terror



Selmer Bringsjord
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
for *Minds & Machines* @ RPI
111507_1930NY

Rensselaer | DEPARTMENT OF COGNITIVE SCIENCE

Rensselaer — Computer Science

Only technology triad can tame terror

Selmer Bringsjord

What killed Goliath? If this question spawns normal associationistic thinking on your part, you will visualize the famous Biblical battle, and will thus find my question peculiar. Don't I mean: *Who* killed Goliath? Actually, I don't.

David was a remarkably brave warrior, yes. And if Jesse's shepherd son hadn't come to the front line, Goliath and his fellow Philistines would have continued taunting a cowering Israel. Nonetheless, however gifted the warrior, however courageous and agile and clever the soldier, the fact remains that if he doesn't carry engineered thunder in his hands, the battle will be lost. David had in hand a nifty little techno-wonder able to rocket a stone to a velocity sufficient to blast clean through the thickest of skulls.

So, again: What killed Goliath? Yes, the sling. If you insist on pressing the Who query, I will certainly admit that David did save the day — but in collaboration with the person responsible for engineering a weapon that made a mockery of the big-mouthed giant. Sticks and stones may break my bones, but names will never hurt me. This David knew, and Goliath, in that split second between a sensation on his forehead and loss of consciousness (and soon thereafter, loss of head), learned.

In the war on terror, the Occidental world will lose, unless its brilliant engineers are suitably funded and tasked. We can increase the number of soldiers a thousandfold, or even ten thousandfold, but if they don't bring the lethal sting provided by the engineers back home into battle, death will come across them, and all will be lost.

Even a cursory glance at military history confirms my thesis with a ring of iron. My rather violent ancestors were hard to beat, not just because they were big and fierce, but because Viking weapons and sea transport marked high water marks in the day's technology. Centuries forward in time, nothing has changed: Hitler lost, but he was rather formidable — in large part because his soldiers were armed by brilliant (if morally warped) minds able to put high technology onto the battlefield.

In our case, we sequestered brilliant minds in the Manhattan Project, funded them to the hilt, and in two blinding flashes the Japanese were finished. Soldiers flew the missions, but Enola Gay and Bockscar, and the thunder carried in their bellies, were built by the brains back home. Once those bellies opened, the rest was all engineering. The kamikazes then are like suicide bombers now. The former were vanquished by the engineers, and the latter, if we spend the money on the triad described herein, can be eliminated as well.

In Iraq, the morass will be fixed (and future asymmetrical conflict quickly won) only if our engineers are paid to give us the triad.

Our engineers must be given the resources to produce the perfected marriage of a trio: pervasive, all-seeing sensors; automated reasoners; and autonomous, lethal robots. In short, we need small machines that can see and hear in every corner; machines smart enough to understand and reason over the raw data that these sensing machines perceive; and machines able to instantly and infallibly fire autonomously on the strength of what the reasoning implies.

Concretely, what would the well-funded merger of this trio mean for the war on terror? This: If you are wearing explosives of any kind outside a subterranean environment, you will be spotted by intelligent unmanned airborne sensors, and will be instantly immobilized by a laser or particle beam from overhead. Sensors on and beneath the surface of the Earth will find you, and you will be killed soon thereafter by AI-guided bunker-boring bombs. If you are a murderous dictator like Saddam , a supersonic robot jet no bigger than a dragonfly will take off in the states, thousands of miles from your "impregnable" lair, and streak in a short time directly into your body, depositing a fatal poison like Polonium therein.

If you seek to seize a jetliner with a plan to blow it up or use it as a missile, one biometric scan of your retina before boarding, and lightning-quick reasoning behind the scenes will flag you as a fiend, and you will be quickly greeted by law enforcement, and escorted into a system of interrogation that uses sensors to read secret information directly from your brain: lying will be silly. Want to bring a backpack bomb somewhere, and leave it behind? The contents of your pack will be sensed the second you bring it toward civilization, and it will be vaporized. Interested in the purchase of handguns for Cho-like mayhem? The slightest blip in your background will be discovered in a second, and you will be out of luck. In fact, guns can themselves bear the trio: If you have one, and wish to fire it, it must sense your identity and location and purpose, and run a check to clear the trigger pull — all in a nanosecond. Life-saving examples of the triad in action could be multiplied ad infinitum.

What can lift us from our present course, in which American civilians and soldiers are sitting ducks getting shot down day after day, to the rock-solid safety of the triad? The same thing that lifted us from the road to Hitler-hell that we were on before the Manhattan Project: a well-funded government program in which engineers from the three relevant fields are brought together, and paid to work their lethal magic.

Selmer Bringsjord is professor of cognitive and computer science at Rensselaer Polytechnic Institute and director of the Rensselaer AI & Reasoning Lab there.

Available at:

http://kryten.mm.rpi.edu/tameterror.pdf

So, machines

So, machines

- Sense

So, machines

- Sense
- Reason

So, machines

- Sense
- Reason
- Shoot

A scary future?

A scary future?

Perhaps. But, leaving my recommendation aside for the moment, our future *still* looks scary.

Our Future

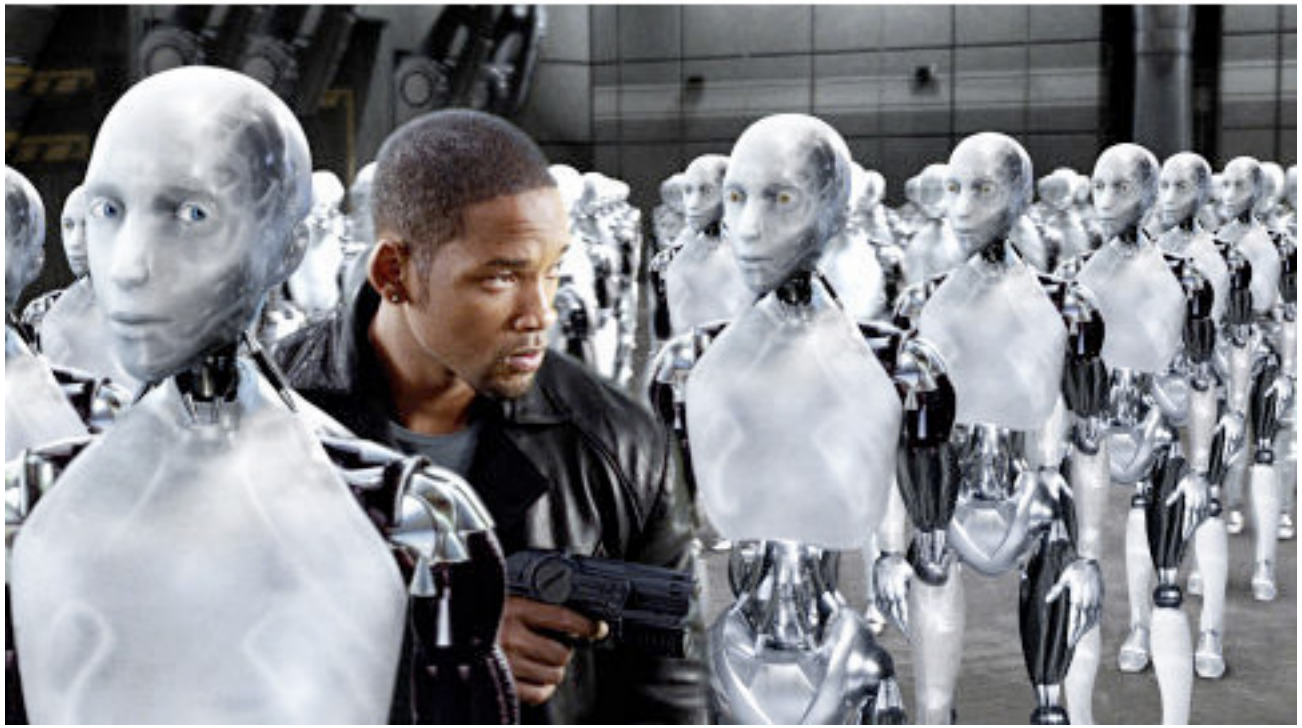
Robots on the battlefield.
Robots in our hospitals.
Robots in law enforcement.
...

Our Problem

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

Our Problem

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



Keynote Presentation

AAAI Symposium on Machine Ethics

Bill Joy in "The Future Doesn't Need Us":

"In the relatively near future, and certainly sooner or later, the human species will be destroyed by advances in robotics technology that we can foresee from our current vantage point, at the start of the new millennium."

I have shown that Joy's argument is unsound in "Ethical Robots: The Future Can Heed Us" in *AI & Society*, a paper based on this presentation. This paper is available in preprint form at:

http://kryten.mm.rpi.edu/sb_aiandsociety_051706_0800.doc

What to *Really* Fear

We need to fear those among us with just enough brain power to use either G or N or R as a weapon.

> Joy: "Thus we have the possibility not just of weapons of mass destruction but of knowledge-enabled mass destruction (KMD), this destructiveness hugely amplified by the power of self-replication. I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals."

Keynote Conclusion

Mosquitoes replicate, as do a thousand thousand other pests. *Ceteris paribus*, robots, whether big or small, at least as I see it, will be at worst pests when left to their own devices. But some humans will no doubt seek to use robots (and for that matter softbots) as weapons against innocent humans. This is undeniable; we can indeed sometimes see the future, and it does look, at least in part, very dark. But it won't be the robots who are to blame. *We* will be to blame. The sooner we stop worrying about inane arguments like those Joy offers, and start to engineer protection against those who would wield robots as future swords, the better off we'll be.

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?

The Solution

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

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 $C = O^\star$

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

This *IEEE Intelligent Systems* paper is available in preprint form at:

http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf

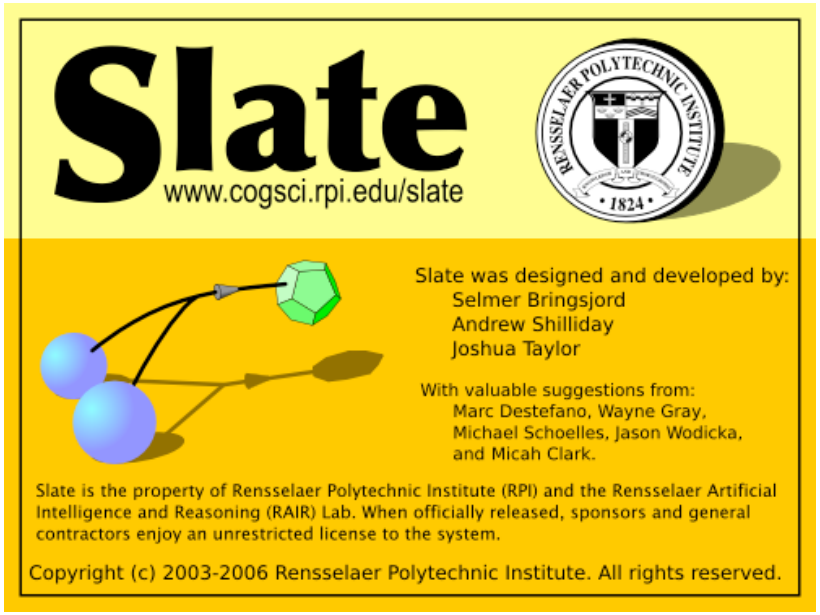
But There is a Twist

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.



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.

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.

Enumerative induction will get us killed.

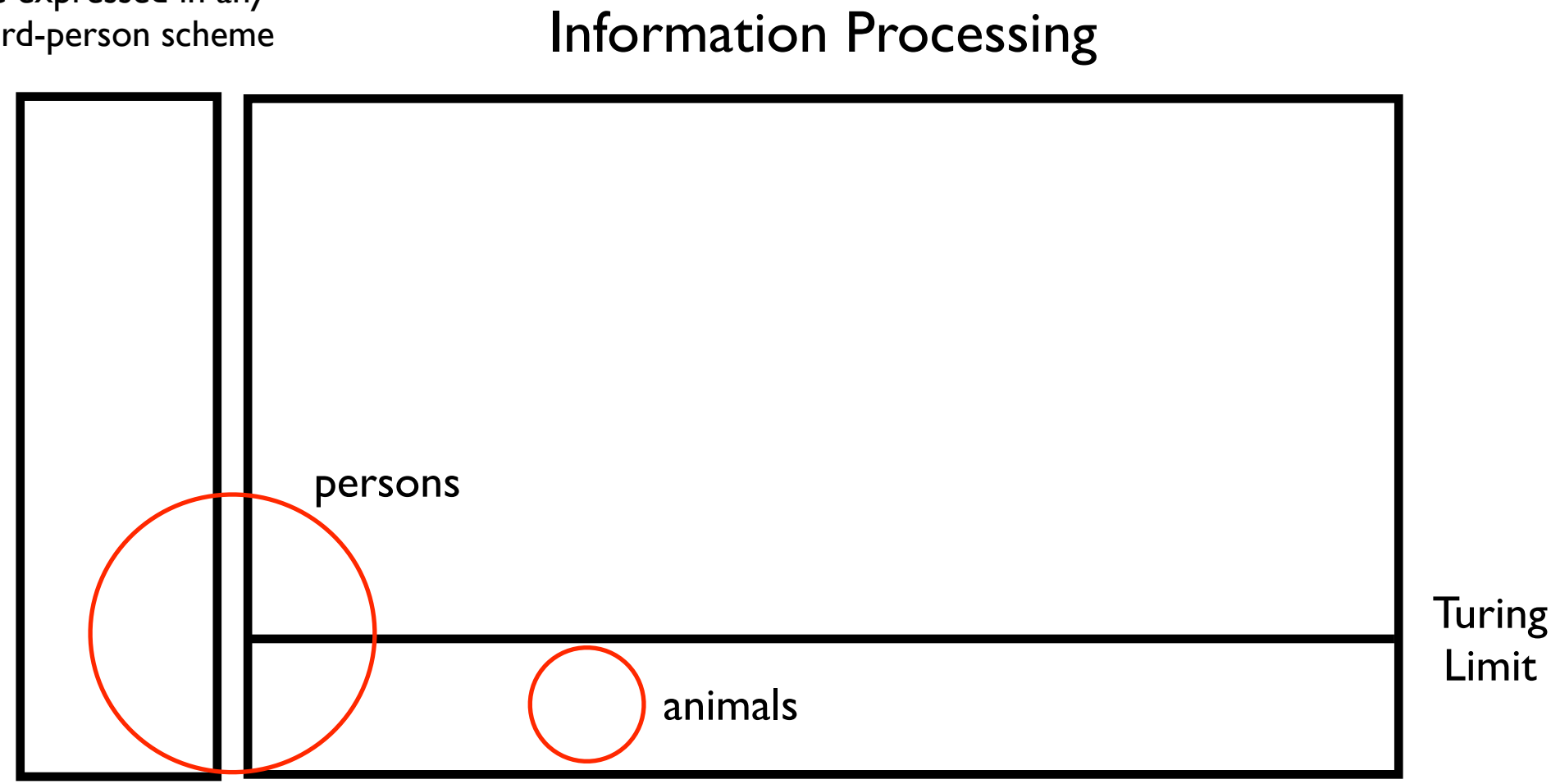
Logic is our only hope, ladies and gentlemen.

And logic will allow the Triad to be built and safely used.

Finis

Superminds (2003)

Phenomena that can't be expressed in any third-person scheme



Superminds (2003)