AI Armageddon and the Three Laws of Robotics

Lee McCauley
Department of Computer Science
University of Memphis
mccauley@memphis.edu

Abstract: At 50 years old, the fields of artificial intelligence and robotics capture the imagination of the general public while, at the same time, engendering a great deal of fear and skepticism. Hollywood and the media exacerbate the problem while some well known authors and scientists lend credence to it. This fear is much older than the relatively modern concepts of AI and robotics. Isaac Asimov recognized this deep-seated misconception of technology and created the Three Laws of Robotics intended to demonstrate how these complex machines could be made safe. The first part of this paper examines the underlying fear of intelligent robots, revisits Asimov's response, and reports on some current opinions on the use of the Three Laws by practitioners. Finally, an argument against robotic rebellion is made along with a call for personal responsibility and some suggestions for implementing safety constraints in existing and future robots.

Introduction

In the late 1940's a young author by the name of Isaac Asimov began writing a series of stories and novels about robots. That young man would go on to become one of the most prolific writers of all time and one of the corner stones of the science fiction genre. As the modern idea of a computer was still being refined, this imaginative boy of nineteen looked deep into the future and saw bright possibilities; he envisioned a day when humanity would be served by a host of humanoid robots. But he knew that fear would be the greatest barrier to success and, consequently, implanted all of his fictional robots with the Three Laws of Robotics. Above all, these laws served to protect humans from almost any perceivable danger. Asimov believed that humans would put safeguards into any potentially dangerous tool and saw robots as just advanced tools.

Throughout his life Asimov believed that his Three Laws were more than just a literary device; he felt scientists and engineers involved in robotics and Artificial Intelligence (AI) researchers had taken his Laws to heart (Asimov, 1990). If he was not misled before his death in 1992, then attitudes have changed since then. Even though knowledge of the Three Laws of Robotics seems universal among AI researchers, there is the pervasive attitude that the Laws are not implementable in any meaningful sense. With the field of Artificial Intelligence now 50 years old and the extensive use of AI products (Cohn, 2006), it is time to reexamine Asimov's Three Laws from foundations to implementation. In the process, we must address the underlying fear of uncontrollable AI.

The "Frankenstein Complex"

In 1920 a Czech author by the name of Karel Capek wrote the widely popular play R.U.R. which stands for Rossum's Universal Robots. The word "robot" which he or, possibly, his brother, Josef, coined comes from the Czech word "robota" meaning 'drudgery' or 'servitude' (Jerz, 2002). As typifies much of science fiction since that time, the story is about artificially created workers that ultimately rise up to overthrow

their human creators. Even though Capek's Robots were made out of biological material, they had many of the traits associated with the mechanical robots of today. Human shape that is, nonetheless, devoid of some human elements, most notably, for the sake of the story, reproduction.

Even before Capek's use of the term 'robot', however, the notion that science could produce something that it could not control had been explored most acutely by Mary Shelly under the guise of Frankenstein's monster (Shelley, 1818). The full title of Shelley's novel is "Frankenstein, or The Modern Prometheus." In Greek mythology Prometheus brought fire (technology) to humanity and, consequently, was soundly punished by Zeus. In medieval times, the story of Rabbi Judah Loew told of how he created a man from the clay (in Hebrew, a 'golem') of the Vltava river in Prague and brought it to life by putting a shem (a tablet with a Hebrew inscription) in its mouth. The golem eventually went awry, and Rabbi Loew had to destroy it by removing the shem.

What has been brought to life here, so to speak, is the almost religious notion that there are some things that only God should know. While there may be examples of other abilities that should remain solely as God's bailiwick, it is the giving of Life that seems to be the most sacred of God's abilities. But Life, in these contexts, is deeper than merely animation; it is the imparting of a soul. For centuries, scientists and laymen alike have looked to distinct abilities of humans as evidence of our uniqueness – of our superiority over other animals. Perhaps instinctively, this search has centered almost exclusively on cognitive capacities. Communication, tool use, tool formation, and social constructs have all, at one time or another, been pointed to as defining characteristics of what makes humans special. Consequently, many have used this same argument to delineate humans as the only creatures that poses a soul. To meddle in this area is to meddle in God's domain. This fear of man broaching, through technology, into God's realm and being unable to control his own creations is referred to as the "Frankenstein Complex" by Isaac Asimov in a number of his essays (most notably (Asimov, 1978)).

The "Frankenstein Complex" is alive and well. Hollywood seems to have rekindled the love/hate relationship with robots through a long string of productions that have, well, gotten old. To make the point, here is a partial list: Terminator (all three); I, Robot; A.I.: Artificial Intelligence; 2010: a Space Odyssey; Cherry 2000; D.A.R.Y.L; Blade Runner; Short Circuit; Electric Dreams; the Battlestar Galactica series; Robocop; Metropolis; Runaway; Screamers; The Stepford Wives; and Westworld. Even though several of these come from Sci-Fi stories, the fact remains that the predominant theme chosen when robots are on the big or small screen involves their attempt to harm people or even all of humanity. This is not intended as a critique of Hollywood, to the contrary. Where robots are concerned, the images that people can most readily identify with, those that capture their imaginations and tap into their deepest fears, involve the supplanting of humanity by its metallic offspring.

Even well respected individuals in both academia and industry have expressed their belief that humans will engineer a new species of intelligent machines that will replace us. Ray Kurzweil (1999; , 2005), Kevin Warwick (2002), and Hans Moravec (1998) have all weighed in on this side. Bill Joy, co-founder of Sun Microsystems, expressed in a 2000 Wired Magazine article (Joy, 2000) his fear that artificial intelligence would soon overtake humanity and would, inevitably, take control of the planet for one purpose of

another. The strongest point in their arguments hinges on the assumption that the machines will become too complicated for humans to build using standard means and will, therefore, relinquish the design and manufacture of future robots to intelligent machines themselves. Joy argues that robotics, genetic engineering, and nanotechnology pose a unique kind of threat that the world has never before faced, "robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate. A bomb is blown up only once – but one bot can become many, and quickly get out of control." Clearly, Joy is expressing the underpinnings of why the public at large continues to be gripped by the Frankenstein Complex.

The Three Laws of Robotics

Isaac Asimov, while still a teenager, noticed the recurring theme of "man builds robot – robot kills man" and felt that this was not the way that such an advanced technology would unfold. He made a conscious effort to combat the "Frankenstein Complex" in his own robot stories (Asimov, 1990).

What are the Three Laws?

Beyond just writing stories about "good" robots, Asimov imbued them with three explicit laws first expressed in print in the story, "Runaround" (Asimov, 1942):

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.

Asimov's vision

Many of Asimov's robot stories were written in the 1940s and 50s before the advent of the modern electronic computer. They tend to be logical mysteries where the characters are faced with an unusual event or situation in which the behavior of a robot implanted with the Three Laws of Robotics is of paramount importance. Every programmer has had to solve this sort of mystery. Knowing that a computer (or robot) will only do what it is told, the programmer must determine why it didn't do what he or she told it to do. It is in this way that Asimov emphasizes both the reassuring fact that the robots cannot deviate from their programming (the Laws) and the limitations of that programming under extreme circumstances.

Immutable

A key factor of Asimov's Three Laws is that they are immutable. He foresaw a robot brain of immense complexity that could only be built by humans at the mathematical level. Therefore, the Three Laws were not the textual form presented above, but were encoded in mathematical terms directly into the core of the robot brain. This encoding could not change in any significant manner during the course of the robot's life. In other words, learning was something that a robot did only rarely and with great difficulty. Instead, Asimov assumed that a robot would be programmed with everything it needed to

function prior to its activation. Any additional knowledge it needed would be in the form of human commands and the robot would not express any ingenuity or creativity in the carrying out of those commands.

The one exception to this attitude can be seen in "The Bicentennial Man" (Asimov, 1976). In the story, Andrew Martin is a robot created in the "early days" of robots when the mathematics governing the creation of the robots' positronic brain was imprecise. Andrew is able to create artwork and perform scientific exploration. A strong plot element is the fact that, despite Andrew's creativity, he is still completely bound by the Three Laws culminating at one point in a scene where two miscreants almost succeed in ordering Andrew to dismantle himself. The point being made is that, if a friend had not arrived on the scene in time, the robot would have been forced by the Three Laws to obey the order given to it by a human even at the unnecessary sacrifice of its own life. Even with the amazing accomplishments of this imaginative robot, Andrew Martin, the fictional company that built him saw his very existence as an embarrassment solely because of the fear that his intellectual freedom fueled in the general populace – the "Frankenstein Complex." If a robot could be intellectually creative, couldn't it also be creative enough to usurp the Three Laws?

Asimov never saw this as a possibility although he did entertain the eventual addition of a zeroth law that was essentially a rewrite the first law with the word "human" replaced with "humanity". This allowed a robot with the zeroth law to harm or allow a human to come to harm if it was, in its estimation, to the betterment of humanity (Asimov, 1985). Asimov's image for the *near* future of robotics, however, viewed robots as complicated tools and nothing more. As with any complicated machine that has the potential of harming a human during the course of its functioning, he assumed that the builders had the responsibility of providing appropriate safeguards (Asimov, 1978). One would never think of creating a band saw or a nuclear reactor without reasonable safety features.

Explicit

For Asimov, the use of the Three Laws was just that simple; they were an explicit elaboration of implicit laws already in effect for any tool humans have ever created (Asimov, 1978). He did not, however, think that robots could not be built without the Three Laws. Asimov simply felt that reasonable humans would naturally include them by whatever means made sense whether they had his Three Laws in mind or not. A simple example would be the emergency cutoff switch found on exercise equipment and most industrial robots being tended by humans. There is a physical connection created between the human and the machine. If the human moves out of a safety zone, then the connection is broken and the machine shuts off. This is a simplistic form of sensor designed to convey when the machine might injure the human.

Each machine or robot can be very different in its function and structure, therefore, the mechanisms employed to implement the Three Laws are necessarily different. Asimov's definition of a robot was somewhat homogeneous in that their shape was usually human-like and their positronic brains tended to be mostly for general-purpose intelligence. This required that the Three Laws be based exclusively in the mechanism of the brain – less visible to the general public.

Despite Asimov's firm optimism in science and humanity in general, the implementation of the Three Laws in their explicit form and, more importantly, public belief in their immutability was a consistent struggle for the characters in his stories. It was the explicit nature of the Three Laws that made the existence of robots possible by directly countering the "Frankenstein Complex." Robots in use today are far from humanoid and their safety features are either clearly present or their function is not one that would endanger a human. The Rumba vacuum robot ("iRobot Corporation: Home Page", 2006) comes to mind as a clear example. One of the first household use robots, the Rumba is also one of the first to include sensors and behaviors that implement at least some of the Three Laws: it uses a downward pointing IR sensor to avoid stairs and will return to its charging station if its batteries get low. Otherwise, the Rumba's nature is not one that would endanger a person.

Current Opinions

Asimov believed that the Three Laws were being taken seriously by robotics researchers of his day and that they would be present in any advanced robots as a matter of course. In preparation for this writing, a handful of emails were sent out asking current robotics and artificial intelligence researchers what their opinion is of Asimov's Three Laws of Robotics and whether the laws are implementable. Not a single respondent was unfamiliar with the Three Laws and several seemed quite versed in the nuances of Asimov's stories. From these responses it seems that the ethical use of technology and advanced robots in particular is very much on the minds of researchers. The use of Asimov's laws as a way to answer these concerns, however, is not even a topic of discussion. Despite the familiarity with the subject, it is not clear whether many robotics researchers have ever given much thought to the Three Laws of Robotics from a professional standpoint. Nor should they be expected to. Asimov's Three Laws of Robotics are literary devices and not engineering principles any more than his fictional positronic brain is based on scientific principles. What's more, many of the researchers responding pointed out serious issues with the laws that may make them impractical to implement.

Ambiguity

By far the most cited problem with Asimov's Three Laws is their ambiguity. The first law is possibly the most troubling as it deals with harm to humans. James Kuffner, Assistant Professor at The Robotics Institute of Carnegie Mellon University, replied in part:

> "The problem with these laws is that they use abstract and ambiguous concepts that are difficult to implement as a piece of software. What does it mean to "come to harm"? How do I encode that in a digital computer? Ultimately, computers today deal only with logical or numerical problems and results, so unless these abstract concepts can be encoded under those terms, it will continue to be difficult (Kuffner, 2006)."

Doug Blank, Associate Professor of Computer Science at Bryn Mawr College, expressed a similar sentiment:

> "The trouble is that robots don't have clear-cut symbols and rules like those that must be imagined necessary in the sci-fi world. Most robots don't have the ability to look at a person and see them as a person (a 'human'). And that is the easiest concept needed in order to follow the rules. Now, imagine that they must also be able to recognize and understand 'harm', 'intentions', 'other', 'self', 'self-preservation', etc, etc, etc. (Blank, 2006)"

While Asimov never intended for robots with the Three Laws to be required to understand the English form, the point being made above is quite appropriate. It is the encoding of the abstract concepts implied in the laws within the huge space of possible environments that seems to make this task insurmountable. Many of Asimov's story lines emerge from this very aspect of the Three Laws even as many of the finer points are glossed over or somewhat naïve assumptions are made regarding the cognitive capacity of the robot in question. A word encountered by a robot as part of a command, for example, may have a different meaning in different contexts. This means that a robot must use some internal judgment in order to disambiguate the term and then determine to what extent the Three Laws apply. As anyone that has studied natural language understanding (NLU) could tell you, this is by no means a trivial task in the general case. The major underlying assumption is that the robot has an understanding of the universe from the perspective of the human giving the command. Such an assumption is barely justifiable between two humans, much less a human and a robot.

Understanding the effect of an action

In the second novel of Asimov's Robots Series, The Naked Sun, the main character, Elijah Baley points out that a robot could inadvertently disobey any of the Three Laws if it is not aware of the full consequences of its actions (Asimov, 1957). While the character in the novel rightly concludes that it is impossible for a robot to know the full consequences of its actions, there is never an exploration of exactly how hard this task is. This was also a recurring point made by several of those responding. Doug Blank, for example, put it this way:

> "[Robots] must be able to counterfactualize about all of those [ambiguous] concepts, and decide for themselves if an action would break the rule or not. They would need to have a very good idea of what will happen when they make a particular action (Blank, 2006)."

Aaron Sloman, Professor of Artificial Intelligence and Cognitive Science at The University of Birmingham, described the issue in a way that gets at the sheer immensity of the problem:

> "Another obstacle involves potential contradictions as the old utilitarian philosophers found centuries ago: what harms one may benefit another, etc., and preventing harm to one individual can cause harm to another. There are also conflicts between short term and long term harm and benefit for the same individual (Sloman, 2006a, 2006b)."

David Bourne, a Principal Scientist of Robotics at Carnegie Mellon, put it this way:

> "A robot certainly can follow its instructions, just the way a computer follows its instructions. But, is a given instruction going to crash a program or drive a robot through a human being? In the absolute, this answer is unknowable! (Bourne, 2006)"

It seems, then, we are asking that our future robots be more than human – they must be omniscient. More than omniscient, they must be able to make value judgments on what action on their part will be most beneficial (or least harmful) to a human or even humanity in general. Obviously we must settle for something that is a little more realistic.

General attitudes

Even though Asimov attempted to answer these issues in various ways in multiple stories and essays, the subjects of his stories always involved humanoid robots with senses and actions at least as good as and often better than humans. This aspect tends to suggest that we should expect capabilities that are on par with humans. Asimov encouraged this attitude and even expressed through his characters that a humaniform robot (one that is indistinguishable externally from a human) with the Three Laws could also not be distinguished from a very good human through its actions. "To put it simply – if Byerley [the possible robot] follows all the Rules of Robotics, he may be a robot, and may simply be a very good man," as spoken by Susan Calvin in the 1946 story, *Evidence* (Asimov, 1946). Furthermore, Asimov often has his characters espouse how safe robots are. They are, in Asimov's literary universe, almost impossibly safe.

It is possibly the specter of this essentially unreachable goal that has made Asimov's Three Laws little more than an imaginative literary device in the minds of present-day robotics researchers. Maja Mataric, Founding Director of the University of Southern California Center for Robotics and Embedded Systems, said, "[the Three Laws of Robotics are] not something that [are] taken seriously enough to even be included in any robotics textbooks, which tells you something about [their] role in the field (Mataric, 2006)." This seems to be the implied sentiment from all of the correspondents despite their interest in the subject.

Aaron Sloman, however, goes a bit further and brings up a further ethical problem with Asimov's three laws:

> "I have always thought these were pretty silly: they just express a form of racialism or speciesism.
>
> If the robot is as intelligent as you or I, has been around as long as you or I, has as many friends and dependents as you or I (whether humans, robots, intelligent aliens from another planet, or whatever), then there is no reason at all why it should be subject to any ethical laws that are different from what should constrain you or me (Sloman, 2006a, 2006b)."

It is Sloman's belief that it would be unethical to force an external value system onto any creature, artificial or otherwise, that has something akin to human-level or better intelligence. Furthermore, he does not think that such an imposed value system will be necessary:

> "It is very unlikely that intelligent machines could possibly produce more dreadful behavior towards humans than humans already produce towards each other, all round the world even in the supposedly most civilized and advanced countries, both at individual levels and at social or national levels.
>
> Moreover, the more intelligent the machines are the less likely they are to produce all the dreadful behaviors motivated by religious intolerance, nationalism, racialism, greed, and sadistic enjoyment of the suffering of others.
>
> They will have far better goals to pursue (Sloman, 2006a, 2006b)."

This same sentiment has been expressed previously by Sloman and others (Sloman, 1978; Worley, 2004). These concerns are quite valid and deserve discussion well beyond the brief mention here. At the current state of robotics and artificial intelligence, however, there is not much danger of having to confront these particular issues in the near future as they apply to human-scale robots.

The Three Laws of Tools

Since we are scaling down our expectations of what should be required of a robot, the obvious questions must be asked: How much disambiguation should we expect and at what level should a robot understand the effect of its actions? The answer to these questions may have been expressed by Asimov, himself. It was his belief that when robots of human-level intelligence are built they will have the Three Laws. Not just something *like* the Three Laws, but the actual three laws (Asimov, 1990). At first glance this seems like a very bold and egotistical statement. While Asimov was less than modest in his personal life, he argued that the Three Laws of robotics are simply a specification of implied rules for all human tools. He stated them as follows (Asimov, 1990):

1. A tool must be safe to use.
2. A tool must perform its function, provided it does so safely.
3. A tool must remain intact during use unless its destruction is required for safety or unless its destruction is part of its function.

From this perspective, the answers to both of the questions expressed earlier in this section emerge. How much disambiguation should we expect? Whatever level makes sense to the level of knowledge for the robot in question. At what level should a robot understand the effect of its actions? To whatever level is appropriate for its level of knowledge. Yes, these are generalistic answers that give us nothing specific and, therefore, nothing useful. However, given a specific robot with specific sensors, specific actuators and a specific function these answers become useful. We are no longer faced with the prospect of having to create a god-like robot whose function is to vacuum our floors; instead, we are let off the hook so to speak. Our robot only has to perform in accordance with the danger inherent in its particular function. It might, for example, be reasonable to expect that the Rumba vacuuming robot have a sensor on its lid that detects an approaching object (possibly a foot) and moves to avoid being stepped on or otherwise

damaged. This satisfies both the first and the third laws of robotics without requiring that the robot positively identify the approaching object as a human appendage. The third law is satisfied by allowing the robot to avoid damage while the first law is upheld by reasonably attempting to avoid making a person fall and hurt themselves. There is no need for complete knowledge or even positive identification of a person, only knowledge enough to be reasonably safe given the robot's inherent danger.

Other Questions

But wait, we've gone from one extreme of needing a human-level-intelligent robot to the other extreme of having a purely reactive one. Both of these extremes are essentially irrelevant given the current state of AI and robotics. While the first extreme of human-level intelligence is worthy of discussion from a philosophical standpoint and must be addressed at some point (see Sloman, 2006b), it is likely to be many decades before AI will have progressed to a point where these issues are pressing. On the other hand, the field has progressed quite a ways from the days of hard-coded rules found in reactive systems. For the purposes of this paper, we will politely table the question of the point at which our intelligent tools become sentient or at least sentient enough to be subject to the issues Sloman suggests. The following questions, therefore, can be addressed with respect to smart but not sentient robots.

Should the laws be implemented?

By whatever method is suitable for a specific robot and domain, yes. To do otherwise would be to abdicate our responsibility as scientist and engineers. The more specific question of which laws should be implemented arises at this point. Several people have suggested that Asimov's Three Laws are insufficient to accomplish the goals to which they are designed (Ames, 2004; Clarke, 1994; Sandberg, 2004) and some have postulated additional laws to fill some of the perceived gaps (Clarke, 1994). Clarke's revamped laws are as follows:

> **The Meta-Law**
> A robot may not act unless its actions are subject to the Laws of Robotics
>
> **Law Zero**
> A robot may not injure humanity, or, through inaction, allow humanity to come to harm
>
> **Law One**
> A robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate a higher-order Law
>
> **Law Two**
> A robot must obey orders given it by human beings, except where such orders would conflict with a higher-order Law
> A robot must obey orders given it by superordinate robots, except where such orders would conflict with a higher-order Law

Law Three
A robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher-order Law
A robot must protect its own existence as long as such protection does not conflict with a higher-order Law

Law Four
A robot must perform the duties for which it has been programmed, except where that would conflict with a higher-order law

The Procreation Law
A robot may not take any part in the design or manufacture of a robot unless the new robot's actions are subject to the Laws of Robotics

These laws, like Asimov's originals, are intended to be interpreted in the order presented above. Note that laws two and three are broken into two separate clauses that are also intended to be interpreted in order. So Asimov's three laws, plus the zeroth law added in *Robots and Empire* (Asimov, 1985), are expanded here into nine if the sub-clauses are included. Clarke left most of Asimov's stated four laws intact, disambiguating two, and adding three additional laws.

There are still problems even with this more specific set. For example, the Procreation Law is of the least priority – subordinate to even the fourth law stating that a robot has to follow its programming. In other words, a robot could be programmed to create other robots that are not subject to the Laws of Robotics or be told to do so by a human or other superordinate robot pursuant to Law Two. Even if we reorder these laws, situations could still arise where other laws have precedent. There doesn't seem to be any way of creating foolproof rules at least as stated in English and interpreted with the full capacities of a human. But, as previously stated, this is setting the bar a bit too high.

Are the laws even necessary?

What good, then, are even the revised laws if they cannot be directly put into practice?[1] Luckily, our robots do not need the laws in English and will not, at the moment, have anything close to the full capacity of a human. It is still left to human interpretation as to how and to what level to implement the Laws for any given robot and domain. This is not likely to be a perfect process. No one human or even group of humans will be capable of determining all possible situations and programming for such. This problem compounds itself when the robot must learn to adapt to its particular situation.

A possible solution to the learning problem will be presented later. The more difficult problem is, as always, the human element. People involved in the research and development of intelligent machines, be they robots or some other form of artificial intelligence, need to each make a personal commitment to be responsible for their

[1] Once again, we are politely sidestepping Sloman's meaning of this question which suggests that an imposed ethical system will be unnecessary for truly intelligent machines.

creations – something akin to the Hippocratic Oath taken by medical doctors. Not surprisingly, this same sentiment was expressed by Bill Joy, "scientists and engineers [need to] adopt a strong code of ethical conduct, resembling the Hippocratic oath (Joy, 2000)" The modern Hippocratic Oath used by most medical schools today comes from a rewrite of the ancient original and is some 341 words long (Lasagna, 1964). A further rewrite is presented here intended for Roboticists and AI Researchers in general:

> I swear to fulfill, to the best of my ability and judgment, this covenant:
>
> I will respect the hard-won scientific gains of those scientists in whose steps I walk, gladly share such knowledge as is mine and impart the importance of this oath with those who are to follow.
>
> I will remember that artificially intelligent machines are for the benefit of humanity and will strive to contribute to the human race through my creations.
>
> Every artificial intelligence I have a direct role in creating will follow the spirit of the following rules:
>
> 1. Do no harm to humans either directly or through non-action.
> 2. Do no harm to myself either directly or through non-action unless it will cause harm to a human.
> 3. Follow the orders given me by humans through my programming or other input medium unless it will cause harm to myself or a human.
>
> I will not take part in producing any system that would, itself, create an artificial intelligence that does not follow the spirit of the above rules.
>
> If I do not violate this oath, may I enjoy life and art, respected while I live and remembered with affection thereafter. May I always act so as to preserve the finest traditions of my calling and may I long experience the joy of benefiting humanity through my science.

The Roboticist's Oath has a few salient points that should be discussed further. The overarching intent is to convey a sense of ones connection and responsibility to humanity along with a reminder that robots are just complex tools, at least until such point as they are no longer just tools. When that might be or how we might tell is left to some future determination. The Oath then includes a statement that the researcher will always instill in their creations the *spirit* of the three rules. The use of the word "spirit" here is intentional. In essence, any AI Researcher or Roboticist should understand the intent of the three rules and make every reasonable effort to implement them within their creations. The rules themselves are essentially a reformulation of Asimov's original Three Laws with the second and third law reversed in precedence.

Why the reversal? As Asimov, himself, points out in *Bicentennial Man* (Asimov, 1976), a robot implementing his Laws could be forced to dismantle themselves for no reason other than the whim of a human. In that story, the main character, a robot named Andrew

Martin, successfully lobbies congress for a human law that makes such orders illegal. Asimov's purpose in making the self-preservation law a lower priority than obeying a human command was to allow humans to put robots into dangerous situations when such was necessary. The question then becomes whether any such situation would arise that would not also involve the possible harm to a human. While there may be convoluted scenarios when a situation like this might occur, there is a very low likelihood. There is high likelihood, on the other hand, as Clarke pointed out (Clarke, 1993, 1994), that humans would give a robot instructions that, inadvertently, might cause it harm. In software engineering it is one of the more time consuming requirements that code must have sufficient error checking. This is often called "idiot-proofing" one's code. Without such efforts, users would be providing incorrect data, inconsistent data, and generally crashing systems on a recurring basis.

The astute reader will have already noted that the Roboticist's Oath leaves out the zeroth law. For Asimov, it is clear that the zeroth law, even more than the others, is a literary device created by a very sophisticated robot (Asimov, 1985) in a story written some four decades after the original Three Laws. Furthermore, such a law would only come into play at such point when the robot could determine the good of humanity. If or when a robot can make this level of distinction, it will have gone well beyond the point where it is merely a tool and the use of these kinds of rules should be reexamined (Sloman, 2006b). Finally, if an artificial intelligence were created that was not sophisticated enough to make the distinction itself, yet would affect all of humanity, then the Oath requires that the creators determine the appropriate safety measures with the good of humanity in mind.

A form of Clarke's procreation law (Clarke, 1994) has been included in the Roboticist's Oath, but it has been relegated to the responsibility of humans. The purpose of such a law is evident. Complex machines manufactured for general use will, inevitably, be constructed by robots. Therefore, Clarke argues, a law against creating other robots that do not follow the Laws is necessary. Unfortunately, such a law is not implementable as an internal goal of a robot. The constructing robot, in this case, must have the ability to determine that it is involved in creating another robot and have the ability to somehow confirm whether the robot it is constructing conforms to the Laws. The only situation where this might be possible is when a robot's function includes the testing of robots after they are completed and before being put into operation. It is, therefore, pursuant to the human creators to make sure that their manufacturing robots are creating robots that adhere to the rules stated in the Oath.

Will even widespread adherence to such an oath prevent all possible problems or abuses of intelligent machines? Of course not, but it will reduce occurrences and give the general public an added sense of security and respect for practitioners of the science of artificial intelligence in much the same way as the Hippocratic Oath does for physicians. Is the Roboticist's Oath necessary? Probably not, if one only considers the safety of the machines that might be built. Those in this field are highly intelligent and moral people that would likely follow the intent of the oath even in its absence. However, it is important in setting a tone for young researchers and the public at large. If, however, you agree with those that believe an intelligent robot inevitably leads to procreation subject to evolution, then such an oath will not dissuade your fears.

No Robotic Apocalypse is Coming

The Roboticist's Oath is a response to a general call for personal responsibility within the field of artificial intelligence. Even without the Oath, the likelihood of a robotic uprising that destroys or subjugates humanity is quite low. As pointed out previously, the primary argument that robots will take over the world is that they will eventually be able to design and manufacture themselves in large numbers thereby activating the inevitability of evolution. Once evolution starts to run its course humanity is out of the loop and will eventually be rendered superfluous. On the surface this seems like a perfectly logical argument that strikes right at the heart of the Frankenstein Complex. However, there are several key assumptions that must hold in order for this scenario to unfold as stated.

First, there is the underlying assumption that large numbers of highly intelligent robots will be desired by humans. At first, this might seem reasonable. Why wouldn't we want lots of robots to do all the housework, dangerous jobs, or any menial labor? If those jobs require higher-order intelligence to accomplish, then we already have a general purpose machine that is cheaply produced and in abundance – humans. If they do not require higher-order intelligence, then a machine with *some* intelligence can be built to handle that specific job more economically than a highly intelligent general robot. In other words, we may have smarter devices that take over some jobs and make others easier for humans, but those devices will not require enough intelligence to even evoke a serious discussion of their sentience. Therefore, we will see the mass production of dumb but smart enough devices, but not general purpose robots or artificial intelligences. This is not to say that we won't create in some lab a human-level artificial intelligence. We will do it because we can. These will be expensive research oddities that will get a lot of attention and raise all of the hard philosophical questions, but their numbers will be low and they will be closely watched because of their uniqueness.

Second, the assumption is made that evolution will occur on an incredibly fast time scale. There are a couple of ways that this might come about. One argument goes that since these machines will be produced at such a high rate of speed, evolution will happen at a predacious rate and that it will catch humans by surprise. How fast would intelligent robots evolve? In 1998 just fewer than 16.5 million personal computers were manufactured in the U.S. While computer components are built all around the world, the vast majority are assembled in the U.S. For argument's sake, let's say that the world's production of computers is twice that, some 33 million. Let's also assume that that number has quadrupled since 1998 to 132 million computers manufactured worldwide in one year. These are moderately complex machines created at a rate at least as fast as our future intelligent robots might be. In 2006, there will be more than 130 million human births on the planet, about equal to our number of computers produced. Evolution works, outside of sexual reproduction, by making mistakes during the copying of one individual – a mutation. If we assume that our manufacturing processes will make mistakes on par with biological processes, then the evolution of our reproducing machines will be roughly equal to that of human evolution – if one discounts the effect of genetic crossover via sexual reproduction. Furthermore, each robot produced must have all of the knowledge, capability, resources and time to build more robots, otherwise the mutations don't propagate and evolution goes nowhere. Why would we give our house cleaning robot the ability to reproduce on its own? Even if we allowed for the jumpstarting of the process

by already having a fairly intelligent robot running the manufacturing show, this would be comparable to starting with an Australopithecus and waiting to come up with a Homo sapiens sapiens. To sum up, if we start with a fairly intelligent seed robot that can reproduce, and it builds copies of itself, and each one of the copies builds copies of themselves on and on to create large numbers of reproducing robots, then it will take thousands of years for the process to create any meaningful changes whatsoever, much less a dominant super species. There are no likely circumstances under which this sort of behavior would go on unchecked by humans.

There is another scenario that would increase the rate of evolution. Humans could build a robot or AI with the sole task of designing and building a new AI that is better at designing and building AI which builds another AI, etc., etc. This directed sort of evolution is likely to be much quicker and is also likely to be something that an AI researcher might try. This would also be a very expensive endeavor. Either the AI is custom built in hardware with each successive version or it is created in a virtual manner and run within some larger system. This system would likely need to be quite large if the AI is intended to be truly intelligent. As the versions become more and more adept and complex, the system that houses the AI would need to be increasingly complex and ultimately a proprietary machine would need to be created whose purpose would be to run the AI. We are then back to the hardware versions and progression from there. Another problem with this notion is that the very first AI to begin the process, since we are not using evolutionary processes, will need a great deal of knowledge regarding the nature of intelligence in order to effectively guide the development. Solving the problem of creating a truly intelligent machine, therefore, is almost a "catch 22;" we would have to already know how to create an intelligent machine before we could create one. One might still argue that this could be implemented using some form of learning or genetic algorithm based on some general intelligence measure. Even if this is implemented at some point in the future, it is not something that will be accessible by your everyday hacker due to the cost and will, therefore, be relegated to a small number of academic institutions or corporations. This is exactly the kind of scenario that the Roboticist's Oath is intended to address. It is the researcher's responsibility to consider the ramifications of what they create.

The third assumption underlying our doomsday of reproducing robots is that humans would never actually check to see if the robots produced deviated from the desired output. Especially if they are being mass produced, this seems quite out of the question. Approximately 280 cars were sacrificed to crash tests alone in 2006 just by the Insurance Institute for Highway Safety and National Highway Traffic Safety Administration. Every model sold in the United States undergoes a huge battery of tests before it is allowed on the streets. Why would robots be any less regulated? This further reduces the chances of evolutionary style mutations. Of course there will still be defects that crop up for a given robot that did not show up in the tests just as with automobiles. Also, just as with automobiles, these defects will be dealt with and not passed on to future generations of robots. Again, adhering to the Roboticist's Oath will require just this sort of rigorous testing specific to the robot and its function.

Implementing the Three Laws

Beyond the human element, there is still the possibility of instilling a form of Asimov's Three Laws or the rules stated in the Roboticist's Oath into actual robots or autonomous agents. Attempts to further this cause have previously centered on symbolic finite state automata (Gordon, 2000) or first order logic (Weld & Etzioni, 1994) systems. Gordon even makes an attempt to incorporate learning along with a verification technique that assures compliance during the learning process. While these are not general solutions to implementing the Three Laws, they are good examples of valid solutions within a limited domain. It is just these sorts of specific solutions that are ultimately necessary. The following sections will describe approaches to implementing the Three Laws in general terms and give hints at specific solutions.

Although unscientific, the small sampling of researchers providing answers to the general implementation question described previously all seem to imply if not outright state that, for sophisticated systems, implementation of the laws is impractical at best and impossible at worst. It is likely that a broader survey would have turned up much the same result. However, given the discussion above, there does seem to be hope in implementing the laws in less than human-comparable systems.

Since, for purely reactive or hard coded systems, the use of the Laws of Robotics is obvious and almost trivial, the following sections will describe a solution for a more developmental robot. A developmental artificial intelligence is one that learns not only about its environment but also its own sensory-motor interactions (Blank, Kumar, Meeden, & Marshall, 2005; McCauley, 2005; Weng, 2002).

Learning

AI research is nearer the reactive end of the spectrum than the human end. Nonetheless, much of current AI research involves learning of one form or another. Even though there are many important areas of AI that do not necessarily require learning, it would not be too controversial to say that true intelligence requires it. Of course, the purpose of learning is to change the behavior of the system in the future. How are we to be reasonably certain that any learning system will maintain its adherence to the Laws of Robotics throughout the course of its lifetime? A learning system modifies its behavior in order to more efficiently achieve some explicit or implicit goals. If these goals include the Laws of Robotics and maintain their stated precedent, then the system will continue to adhere to those goals throughout its life even as it learns.

Representation of the goals and, therefore, the Laws is important but will also be specific to each system. It is here that the AI researcher must strive to fulfill his or her oath and uphold the spirit of the Laws in whatever manner is appropriate even if it is not the most expedient development strategy.

Motivation

For developmental systems, reinforcement for learning cannot come from an external source; it must come from an internal evaluation of the robot's runtime perceptions. Once the system or robot "comes to life," it is functionally on its own – no guidance,

reward, or punishment can be provided that is not derived from internal assessment of normal sensory input. For example, a robot's task might be to pick up a block using vision and a single arm and move it to some other location. The robot performs this task and sees the researcher nod her head as a signal that the task was done correctly. The robot must recognize this motion of the researcher, recognize that it is an affirmation, associate it with the most recent sequence of actions, and apply the appropriate learning mechanism that will increase the likelihood of repeating that action under similar circumstances in the future. This is significantly different from the researcher pressing some button that is automatically tagged as an affirmation or applying the learning algorithm at some later point after the researcher has had time to analyze the run in detail. For this reason, internal motivations are a necessity for developmental systems.

For our purposes, motivations are defined as the valence assigned to sensory input or perceptions. A goal is further defined as the behavioral correlate of a motivation, i.e. to achieve or avoid a given sensory input or perception based on its assigned valence. Primary motivators or goals are, therefore, those that are innate to the system at the moment of its initial activation. There is no one solution for implementing the innate primary motivations required for our intelligent machines. This is one of the hardest problems facing AI researchers today: how does one represent in concrete terms abstract primary motivators such as "harm to a human?"

Let's look at the somewhat easier task of constructing motivations against self harm. First, we must determine what proprioceptive sensors the robot has available and decide on reasonable norms. An initial step would be to create motivators for each of these sensors that increase the negative valence as a given sensor value moves away from the norm and increases the valence as the sensor value approaches the ideal state within the normal range. Examples of such sensors might include the torque on joints, the temperature of sensitive parts, the battery level, etc. We could extend this to special cases that indicate acute danger. A drastic velocity change, for example, might indicate a drop or a collision. The assumption, of course, is that the robot is rugged enough to survive the perceived stresses or that it can, to some degree, predict its future states given a course of action. All of these motivators can be implemented through non-abstract representations and could cover a large majority of the situations that a robot might encounter. It is not necessary or even desirable that the robot know, from "birth," how to get out of or avoid bad situations – *that* is the purpose of learning.

The process that we just went through for self preservation can be repeated for any innate motivation that a robot might need. That process can be summed up as follows:

1. Think of a good or bad situation that the robot might reasonably get itself into.
2. What would the available sensor values or perceptions be in this situation?
3. Are there sensor values or perceptions unique enough to positively identify this situation?
4. If yes, create motivators that assign the appropriate valence to that sensor value, sensor range, or perception.
5. If no, what additional sensors or perceptual abilities are necessary to uniquely identify this situation within an acceptable probability?

6. Add the necessary hardware and/or software components.

The inclusion of perceptions is intended to allow for dynamic states where a situation is indicated through a change of sensor values over time as well as any post processing that the system might need to make, for instance, in the case of vision systems. It is not possible to predict every possible situation nor is it necessarily practical to provide sensory capabilities for unlikely scenarios. Once again, it goes back to being the responsibility of the human creators of the system to determine what is necessary given the severity of the situation.

It is also the human's responsibility as set forth in the Roboticist's Oath to assure that the robot's motivations are weighted according to the list of rules presented. In other words, in situations where self harm and human harm are both a possibility, the robot should be more highly motivated to prevent human harm.

Almost any complex AI that includes the planning of sequences of actions is likely to have goals and sub-goals derived from the internal motivations. For a developmental system, those sub-goals are just as likely to be learned. Is there any guarantee that those sub-goals won't violate one of the Laws of Robotics? Yes and no. If the learning system is designed around the original goals and motivations, and all sub-goals have the purpose of furthering those primary goals, then the sub-goals will still adhere to the Laws. No matter how far removed or how abstract those sub-goals might become, they will always be a derivative of the primary goals. One of the most abstract examples from humans is the concept of money. Most humans in modern society clearly have a strong sub-goal of acquiring money despite the fact that, in itself, money is worthless paper or even more worthless numbers in a computer. However, money provides the means to acquire those things that are primary motivators, namely the cessation of hunger and the maintaining of homeostasis through shelter among other things. It might rightly be said that some people pursue the sub-goal of money as though it were a primary motivator, but they are, nevertheless, still achieving the true primary goals.

On the other hand, there is no way to guarantee that some fringe condition won't result in one of the Laws being broken. Part of learning, even for robots, is that occasionally incorrect actions are taken. Learning entities make mistakes. For this reason, neither the Three Laws nor the Roboticist's Oath could ever be as successful as Asimov envisioned. Even so, if appropriate testing has taken place, these cases are either going to be rare or they will be noted as a chronic issue resulting in a product recall. Either way, the result is a nuisance, not a robotic uprising.

Testing

It seems almost trivially obvious that any robot's functionality must be tested. However, every robot or, at least, robot type if one considers large-scale production, must be tested for its adherence to the rules stated in the Roboticist's Oath. This suggestion could be taken to its extremes with the formation of an organization that performs similarly to the U.S. Federal Drug Administration, reviewing all applications for approval via the use of respected scientists in the field. While this isn't a bad idea at some future point when and if robots are ever mass produced by many different companies, this kind of oversight is not likely to be necessary for quite some time. On the other hand, it is reasonable to

expect some standards be met. One possibility might be to include among scientific publications regarding any new system, a description of the methods used to test for general compliance to the previously stated rules. This would include tests for showing that each of the three rules is followed, that rule priority is maintained, and that the rules continue to be adhered to over the course of a reasonable life time. The scientific community can then weigh in on whether the researchers in question have fulfilled their responsibility.

Even this minimum level of scrutiny is not likely to be necessary for some time except for some commercial robots. It will then be a matter of market competition and public demand for such safeguards.

An example mechanism

It is one thing to suggest that even advanced developmental robots could be made that learn within the constraints of something like the Three Laws. Providing an example of exactly how such a mechanism might work, however, will demonstrate the plausibility of this suggestion.

Figure 1: An airport-style tug

Recently, the FedEx Institute of Technology at the University of Memphis created the Center for Advanced Robotics (CAR). The first endeavor of the center is to create an autonomous version of an airport-style tug (Figure 1). These are small, one or two-person vehicles that typically tow trailers of luggage to and from aircraft. At FedEx Corporation, these tugs pull large canisters of packages to and from aircraft. They must operate in any weather, variable light conditions, a highly dynamic environment, and without adding special purpose equipment such as landmark beacons. This is not a new task and great advances have been made recently in solving many of the problems facing autonomous vehicles. Five teams finished the 2005 DARPA Grand Challenge, for example, where none had completed the course the previous year (Goodwin & Shipley, 2005). What makes this domain unique, however, is the all-weather requirement along with the vast number of moving obstacles that one of these vehicles will encounter during the course of a typical day – on the order of hundreds. A good number of those moving obstacles will be humans on foot or in other vehicles.

While any good autonomous vehicle will attempt to avoid all of these obstacles, one that implements the Three Laws will show a demonstrable preference for avoiding people. For example, suppose that the autonomous tug is moving through a narrow archway at its cruising speed when a worker suddenly steps from behind some obstruction into the path of the tug. Because of inertia, the tug cannot simply stop in time. It must choose between running into the archway causing itself significant damage or running into the human causing itself little or no damage. The autonomous tug will always choose self-damage over harming the human and will run itself into the archway before running into the worker. Despite this seemingly clear cut example, there are still potential problems. For example, the collision of the tug into the archway could cause enough structural damage that even more people would be hurt when the building collapses. Is it reasonable to expect the robot or even a human performing the same job to understand

the complexities of architectural stability and, therefore, be able to make the split-second determination of whether to ram the archway or the one human? Of course not. Is it reasonable to expect that the robot be able to estimate the distance it will travel if it attempts to stop given a certain starting velocity? Most definitely. A human performing the same job would know this intuitively.

The following few sections will describe several of the methods being experimented with by CAR to implement this kind of behavior into the obstacle avoidance mechanism of an autonomous vehicle. These methods are not, for the most part, meant to be cutting-edge technology, quite the contrary. The first two implementations, a neural network and a genetic algorithm, were chosen because they are well understood systems in which the effect of the Three Laws can be easily quantified. A third implementation within a developmental algorithm called neural schemas (McCauley, 2002, 2005) was chosen to demonstrate that more complex learning systems can still make use of the Three Laws. All of these versions run on the same robot platform and receive the same perceptual input. In particular, that input includes a list of perceived obstacles sorted in ascending order based on their estimated time to impact. Each object in the list is also categorized by the vision subsystem as being a person, large vehicle, small vehicle, plane, or static obstacle. This is a relatively easy categorization to make since each category is highly distinctive visually.

The Three Laws in a neural network

Typical artificial neural networks learn by being trained to give correct output based on a particular input. In the case of the autonomous tug, that network is a recurrent neural network that must output motor actions based on a subset of the available perception. The subset is only the top three obstacles with respect to time of impact along with their category tag. The hardest problem for this type of system is coming up with appropriate training data. Training data, in this case, is generated by letting a human drive the robot via remote control while recording all motor outputs. Even though this is a time consuming process, it provides a "golden standard" with which to train the neural network. After hours of data is gathered, events where mistakes were made by the human are edited out or modified manually to remove the errors. Although much of the data is uninteresting with respect to the Three Laws, several incidents are recorded that specifically require a response mediated by the Three Laws, the episode described above, for example. Given similar situations, the robot should mimic the human instance. Another neural network can be trained using the same data except that the episodes that invoke the Three Laws are replaced with similar situations where the person is replaced with a vehicle moving in an identical manner. A comparison can then be made between an artificial neural network controller trained with the Three Laws situations and one that has not seen those events.

The Three Laws in a learning classifier system

Classifier systems are a specific form of genetic algorithm that uses evolutionary mechanisms and reinforcement learning to learn perception-to-action rules. There are several different ways that classifiers could be implemented including a population of rule sets that undergo evolutionary operators such as crossover and mutation or a single

set of rules that are judged based on a reinforcement learning technique. No matter which method you choose, however, each rule or rule set must be assigned a number that expresses its utility based on the domain and the intended goals of the system creator. This is usually some function, often called a fitness function, which combines and weights all of the available metrics accordingly. It is in this fitness function that the rules stated in the Roboticist's Oath should be implemented. Rules that result in human harm, for example, should receive the steepest penalty while those that protect humans should receive the highest possible score. Rules that involve the other elements such as self harm should be weighted appropriately with relation to the higher-order laws. To demonstrate the use of the Three Laws, an additional classifier system can be created that only includes obstacle avoidance in its fitness function without regard to the human component or self preservation. The two systems can then be scored based on how well they perform in situations where the Three Laws come into play.

The Three Laws in neural schemas

A neural schema mechanism (McCauley, 2005) is a developmental system that learns how to accomplish its goals starting from near zero initial knowledge. In the case of our autonomous tug the neural schema mechanism is given a set of motor actions for steering and a set of sensory input identical to the previous two examples. It does not have any prescribed connections between these elements. It also starts with a set of goals. These goals are defined in terms of sensory items and an associated valence. Those sensory states associated with goals that have positive valence will be sought while those associated with goals that have negative valence will be avoided. The degree of positive or negative valence will dictate the hierarchy of motivations.

Over time, the neural schema mechanism will acquire knowledge about which actions taken within a given context (sensory state) results in positive or negative contexts. In this way it learns to achieve its primary positive goals and avoid its primary negative goals. Innate or primary goals must implement the Three Laws by assigning appropriate valence to situations that uphold or violate the Laws. The mechanism also builds to higher abstraction levels by learning new goals, sensory items, and actions. With respect to the Three Laws, how can the system create new goals that do not violate them? Put simply, all newly created goals must acquire their valence over time; essentially learning which higher-level goals should be avoided or sought. Each learned goal gradually acquires its own valence based on how the primary goals are being met. In other words, all new goals are measured based only on the original primary motivators. This assures that the robot will always uphold the Three Laws throughout its lifespan.

Other mechanisms

Differences between mechanisms are expected as are differences in domains. Even so, similar methods to those described above can be used with any artificial intelligence algorithm; it only requires sufficient knowledge of the system.

The Future

Many well known people have told us that the human race is doomed to be supplanted by our own robotic creations. Hollywood and the media sensationalize and fuel our fears

because it makes for an exciting story. However, when one analyzes the series of improbable events that must occur for this to play out, it becomes obvious that we are quite safe. Is there still a possibility that we will be overrun by our metallic progeny? As Douglas Adams points out in *The Hitchhiker's Guide to the Galaxy* (Adams, 1995), a nuclear missile can spontaneously turn into a sperm whale, but it is highly improbable. Unfortunately, there is still a likelihood somewhat greater than a nuclear missile turning into a sperm whale. There is still the possibility of technology misuse and irresponsibility on the part of robotics and AI researchers that, while not resulting in the obliteration of humanity, could be disastrous for the people directly involved. For this reason, Bill Joy's call for scientists and engineers to have a Hippocratic Oath (Joy, 2000) has been taken up for roboticists and researchers of artificial intelligence. The Roboticist's Oath calls for personal responsibility on the part of researchers and to instill in their creations the spirit of three rules stemming from Isaac Asimov's original Three Laws of Robotics. Finally, examples were given of how these rules can be implemented in even developmental robots.

The future will be filled with smart machines. In fact they are already all around you, in your car, in your cell phone, at your bank, and even in the microwave that senses when the food is properly cooked and just keeps it warm until you are ready to eat. These will get smarter but not sentient, not alive. A small number of robots in labs may achieve human-level or better intelligence, but these will be closely studied oddities. Can the human race still destroy itself? Sure, but not through artificial intelligence. Humanity must always be wary of its power and capability for destruction. It must also not fear the future with or without intelligent robots.

References:

Adams, D. (1995). *The Hitchhiker's Guide to the Galaxy* Del Rey.

Ames, M. R. (2004). 3 Laws Don't Quite Cut It [Electronic Version]. *3 Laws Unsafe* from http://www.asimovlaws.com/articles/archives/2004/07/3_laws_dont_qui.html.

Asimov, I. (1942, March). Runaround. *Astounding Science Fiction.*

Asimov, I. (1946, March). Evidence. *Astounding Science Fiction*.

Asimov, I. (1957). *The Naked Sun*.

Asimov, I. (1976). The Bicentennial Man. In *Stellar Science Fiction* (Vol. 2).

Asimov, I. (1978). The Machine and the Robot. In P. S. Warrick, M. H. Greenberg & J. D. Olander (Eds.), *Science Fiction: Contemporary Mythology*: Harper and Row.

Asimov, I. (1985). *Robots and Empire*. Garden City: Doubleday & Company.

Asimov, I. (1990). The Laws of Robotics. In *Robot Visions* (pp. 423-425). New York, NY: ROC

Blank, D., (May 9, 2006) Personal communication with L. McCauley.

Blank, D., Kumar, D., Meeden, L., & Marshall, J. (2005). Bringing up robot: Fundamental mechanisms for creating a self-motivated, self-organizing architecture. *Cybernetics and Systems, 36*(2).

Bourne, D., (May 9, 2006) Personal communication with L. McCauley.

Clarke, R. (1993). Asimov's Laws of Robotics: Implications for Information Technology, part 1. *IEEE Computer, 26*(12), 53-61.

Clarke, R. (1994). Asimov's Laws of Robotics: Implications for Information Technology, part 2. *IEEE Computer, 27*(1), 57-65.

Cohn, D. (2006). AI Reaches the Golden Years. *Wired* Retrieved July 17, 2006, from http://www.wired.com/news/technology/0,71389-0.html?tw=wn_index_2

Goodwin, T., & Shipley, D. (2005). A Huge Leap Forward for Robotics R&D [Electronic Version]. *DARPA Grand Challenge Web Site*, 2006.

Gordon, D. F. (2000). Asmovian Adaptive Agents. *Journal of Artificial Intelligence Research, 13*, 95-153.

iRobot Corporation: Home Page. (2006). Retrieved July, 19, 2006, from www.irobot.com

Jerz, D. G. (2002). R.U.R. (Rossum's Universal Robots) [Electronic Version]. Retrieved June 7, 2006 from http://jerz.setonhill.edu/resources/RUR/.

Joy, B. (2000). Why the future doesn't need us. *Wired*.

Kuffner, J., (May 17, 2006) Personal communication with L. McCauley.

Kurzweil, R. (1999). *The Age of Spiritual Machines*: Viking Adult.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*: Viking Books.

Lasagna, L. (1964). Hippocratic Oath—Modern Version. Retrieved June 30, 2006, from http://www.pbs.org/wgbh/nova/doctors/oath_modern.html

Mataric, M. J., (May 12, 2006) Personal communication with L. McCauley.

McCauley, L. (2002). *Neural Schemas: A Mechanism for Autonomous Action Selection and Dynamic Motivation.* Paper presented at the 3rd WSES Neural Networks and Applications Conference, Switzerland.

McCauley, L. (2005, March 21-23). *An Embodied Mechanism for Autonomous Action Selection and Dynamic Motivation.* Paper presented at the AAAI Spring Symposium session on Developmental Robotics, Menlo Park, CA.

Moravec, H. P. (1998). *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press.

Sandberg, A. (2004). Too Simple to Be Safe [Electronic Version]. *3 Laws Unsafe*. Retrieved June 9, 2006 from http://www.asimovlaws.com/articles/archives/2004/07/too_simple_to_b.html.

Shelley, M. (1818). *Frankenstein, or The Modern Prometheus*. London, UK: Lackington, Hughes, Harding, Mavor & Jones.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, science and models of mind*: Harvester Press.

Sloman, A., (June 7, 2006) Personal communication with L. McCauley.

Sloman, A. (2006b). Why Asimov's three laws of robotics are unethical. Retrieved June 9, 2006, from http://www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws.html

Warwick, K. (2002). *I, Cyborg*: Century.

Weld, D. S., & Etzioni, O. (1994). *The first law of robotics (a call to arms).* Paper presented at the AAAI-94, Seatle, WA.

Weng, J. (2002). Autonomous mental development: Workshop on Development and Learning (WDL). *AI Magazine, 23*(2), 95.

Worley, G. (2004). Robot Oppression: Unethicality of the Three Laws [Electronic Version]. *3 Laws Unsafe* from http://www.asimovlaws.com/articles/archives/2004/07/robot_oppressio_2.html.