The Frankenstein Complex and Asimov's Three Laws

Lee McCauley

FedEx Institute of Technology
University of Memphis
374 Dunn Hall
Memphis, TN 38152
mccauley@memphis.edu

Abstract

Public fear will be the biggest hurdle for intelligent robots to overcome. Understanding society's long-standing fear of self-aware automatons should be a consideration within robotics labs, especially those specializing in fully autonomous humanoid robots. Isaac Asimov anticipated this fear and proposed the Three Laws of Robotics as a way to mollify it somewhat. This paper explores the "Frankenstein Complex" and current opinions from noted robotics researchers regarding the possible implementation of Asimov's Laws. It is clear from these unscientific responses why the Three Laws are impractical from a general sense even though the ethical issues involved are at the forefront of researchers' minds. The onus is, therefore, placed on the roboticists of today and the future to hold themselves to a standard similar to the Hippocratic Oath that preserves the spirit of Asimov's Laws.

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 and 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 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 possess 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 literature, 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. 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 could soon overtake humanity and would, inevitably, take control of the planet for one purpose or another. Even if his logic is a bit flawed, Joy is expressing the underpinnings of why the public at large continues to be gripped by the Frankenstein Complex. Even though the public is fascinated with the current robots they see demonstrated in documentaries, there seems to be a general fear that these robots will become too intelligent. Is it possible that examining the social ramifications of Asimov's Three Laws and their possible implementation in real robots could ameliorate some of this fear?

Current Opinions in the Field

Asimov believed that his "Three Laws of Robotics" 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 (Asimov 1978; Asimov 1990). In preparation for this writing, a handful of emails were sent out asking current robotics and artificial intelligence researchers what their opinion was of Asimov's Three Laws of Robotics and whether the laws could be implemented. 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, after all, 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 2006; Sloman 2006).

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 and often better than humans. This aspect tends to suggest that we should expect actions and 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, just through its actions, be distinguished from a very good human. "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 2006; Sloman 2006).

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 2006; Sloman 2006).

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.

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 scientists 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 (Clarke 1994; Ames 2004; Sandberg 2004) and some have postulated additional laws to fill some of the perceived gaps (Clarke 1994). For example, Asimov's original three laws, plus the zeroth law added in *Robots and Empire* (Asimov 1985), are expanded by Clarke (1993; 1994) 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 to fill what he considered ambiguous gaps that made them impractical to implement.

There are still problems, however, even with this more specific set. For example, the Procreation Law, stating that a robot cannot take part in the creation of another robot not subject to the laws, 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 would still arise where other laws have precedent. There doesn't seem to be any way of creating a foolproof set of 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? 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.

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 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 itself either directly or through non-action unless it will cause harm to a human.
3. Follow the orders given it by humans through its programming or other input medium unless it will cause harm to itself 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 – in other words, relying on human agreement, however flawed, to protect robots. 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 (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. By reversing the priority of these two rules it is being suggested that researchers "idiot-proof" their creations while keeping human safety paramount.

The astute reader will have also 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 2006). 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. Similar such safeguards have successfully protected humanity from inadvertent nuclear missile launch for decades.

A form of Clarke's procreation law (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. A human wanting to circumvent the Laws could do so quite easily. 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.

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. Even though Asimov's Three Laws of Robotics are not being explicitly implemented in today's advanced robots for reasons laid out in this paper, robotics researchers take these ethical issues quite seriously. Of course, there is still the possibility of technology misuse and irresponsibility on the part of robotics and AI researchers that, while not likely to result 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.

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. As our devices and robots get smarter, we must be cognizant of how the general public perceives our contributions to society. Will they fear them or welcome them? The answer is up to us.

References

Ames, M. R. (2004) "3 Laws Don't Quite Cut It." 3 Laws Unsafe **Volume**, DOI:

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

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

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

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

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

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

Blank, D. (2006). Robotics and Asimov's Three Laws (personal communication). L. McCauley.

Bourne, D. (2006). Robotics and Asimov's Three Laws (personal communication). 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.

Jerz, D. G. (2002) "R.U.R. (Rossum's Universal Robots)." **Volume**, DOI:

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

Kuffner, J. (2006). Regarding Isaac Asimov's Three Laws of Robotics (personal communication). 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. (2006). Robotics and Asimov's Three Laws (personal communication). L. McCauley.

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

Sandberg, A. (2004) "Too Simple to Be Safe." 3 Laws Unsafe **Volume**, DOI:

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. (2006). Robotics and Asimov's Three Laws (personal communication). L. McCauley.

Sloman, A. (2006). "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.

Worley, G. (2004) "Robot Oppression: Unethicality of the Three Laws." 3 Laws Unsafe **Volume**, DOI: