The Fourth Law of Robotics - Part I

By Sam Vaknin

The Fourth Law of Robotics - Part I by Sam Vaknin

The movie "I, Robot" is a muddled affair. It relies on shoddy pseudo-science and a general sense of unease that artificial (non-carbon based) intelligent life forms seem to provoke in us. But it goes no deeper than a comic book treatment of the important themes that it broaches. I, Robot is just another - and relatively inferior - entry is a long line of far better movies, such as "Blade Runner" and "Artificial Intelligence".

Sigmund Freud said that we have an uncanny reaction to the inanimate. This is probably because we know that – pretensions and layers of philosophizing aside – we are nothing but recursive, self aware, introspective, conscious machines. Special machines, no doubt, but machines all the same.

Consider the James bond movies. They constitute a decades-spanning gallery of human paranoia. Villains change: communists, neo-Nazis, media moguls. But one kind of villain is a fixture in this psychodrama, in this parade of human phobias: the machine. James Bond always finds himself confronted with hideous, vicious, malicious machines and automata.

It was precisely to counter this wave of unease, even terror, irrational but all-pervasive, that Isaac Asimov, the late Sci-fi writer (and scientist) invented the Three Laws of Robotics:

A robot may not injure a human being or, through inaction, allow a human being to come to harm.
A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
Many have noticed the lack of consistency and, therefore, the inapplicability of these laws when considered together.

First, they are not derived from any coherent worldview or background. To be properly implemented and to avoid their interpretation in a potentially dangerous manner, the robots in which they are embedded must be equipped with reasonably comprehensive models of the physical universe and of human society.

Without such contexts, these laws soon lead to intractable paradoxes (experienced as a

nervous breakdown by one of Asimov's robots). Conflicts are ruinous in automata based on recursive functions (Turing machines), as all robots are. Godel pointed at one such self destructive paradox in the "Principia Mathematica", ostensibly a comprehensive and self consistent logical system. It was enough to discredit the whole magnificent edifice constructed by Russel and Whitehead over a decade.

Some argue against this and say that robots need not be automata in the classical, Church-Turing, sense. That they could act according to heuristic, probabilistic rules of decision making. There are many other types of functions (non-recursive) that can be incorporated in a robot, they remind us.

True, but then, how can one guarantee that the robot's behavior is fully predictable ? How can one be certain that robots will fully and always implement the three laws? Only recursive systems are predictable in principle, though, at times, their complexity makes it impossible.

This article deals with some commonsense, basic problems raised by the Laws. The next article in this series analyses the Laws from a few vantage points: philosophy, artificial intelligence and some systems theories.

An immediate question springs to mind: HOW will a robot identify a human being? Surely, in a future of perfect androids, constructed of organic materials, no superficial, outer scanning will suffice. Structure and composition will not be sufficient differentiating factors.

There are two ways to settle this very practical issue: one is to endow the robot with the ability to conduct a Converse Turing Test (to separate humans from other life forms) - the other is to somehow "barcode" all the robots by implanting some remotely readable signaling device inside them (such as a RFID - Radio Frequency ID chip). Both present additional difficulties.

The second solution will prevent the robot from positively identifying humans. He will be able identify with any certainty robots and only robots (or humans with such implants). This is ignoring, for discussion's sake, defects in manufacturing or loss of the implanted identification tags. And what if a robot were to get rid of its tag? Will this also be classified as a "defect in manufacturing"?

In any case, robots will be forced to make a binary choice. They will be compelled to classify one type of physical entities as robots – and all the others as "non-robots". Will non-robots include monkeys and parrots? Yes, unless the manufacturers equip the robots with digital or optical or molecular representations of the human figure (masculine and feminine) in varying positions (standing, sitting, lying down). Or unless all humans are somehow tagged from birth.

These are cumbersome and repulsive solutions and not very effective ones. No dictionary of human forms and positions is likely to be complete. There will always be the odd physical posture which the robot would find impossible to match to its library. A human disk thrower or swimmer may easily be classified as "non-human" by a robot - and so might amputated invalids.

What about administering a converse Turing Test?

This is even more seriously flawed. It is possible to design a test, which robots will apply to distinguish artificial life forms from humans. But it will have to be non-intrusive and not involve overt and prolonged communication. The alternative is a protracted teletype session, with the human concealed behind a curtain, after which the robot will issue its verdict: the respondent is a human or a robot. This is unthinkable.

Moreover, the application of such a test will "humanize" the robot in many important respects. Human identify other humans because they are human, too. This is called empathy. A robot will have to be somewhat human to recognize another human being, it takes one to know one, the saying (rightly) goes.

Let us assume that by some miraculous way the problem is overcome and robots unfailingly identify humans. The next question pertains to the notion of "injury" (still in the First Law). Is it limited only to physical injury (the elimination of the physical continuity of human tissues or of the normal functioning of the human body)?

Should "injury" in the First Law encompass the no less serious mental, verbal and social injuries (after all, they are all known to have physical side effects which are, at times, no less severe than direct physical "injuries")? Is an insult an "injury"? What about being grossly impolite, or psychologically abusive? Or offending religious sensitivities, being politically incorrect - are these injuries? The bulk of human (and, therefore, inhuman) actions actually offend one human being or another, have the potential to do so, or seem to be doing so.

Consider surgery, driving a car, or investing money in the stock exchange. These "innocuous" acts may end in a coma, an accident, or ruinous financial losses, respectively. Should a robot refuse to obey human instructions which may result in injury to the instruction-givers?

Consider a mountain climber – should a robot refuse to hand him his equipment lest he falls off a cliff in an unsuccessful bid to reach the peak? Should a robot refuse to obey human commands pertaining to the crossing of busy roads or to driving (dangerous) sports cars?

Which level of risk should trigger robotic refusal and even prophylactic intervention? At which stage of the interactive man-machine collaboration should it be activated? Should a robot refuse to fetch a ladder or a rope to someone who intends to commit suicide by hanging himself (that's an easy one)?

Should he ignore an instruction to push his master off a cliff (definitely), help him climb the cliff (less assuredly so), drive him to the cliff (maybe so), help him get into his car in order to drive him to the cliff... Where do the responsibility and obeisance bucks stop?

Whatever the answer, one thing is clear: such a robot must be equipped with more than a rudimentary sense of judgment, with the ability to appraise and analyse complex situations, to predict the future and to base his decisions on very fuzzy algorithms (no programmer can foresee all possible circumstances). To me, such a "robot" sounds

much more dangerous (and humanoid) than any recursive automaton which does NOT include the famous Three Laws.

Moreover, what, exactly, constitutes "inaction"? How can we set apart inaction from failed action or, worse, from an action which failed by design, intentionally? If a human is in danger and the robot tries to save him and fails – how could we determine to what extent it exerted itself and did everything it could?

How much of the responsibility for a robot's inaction or partial action or failed action should be imputed to the manufacturer – and how much to the robot itself? When a robot decides finally to ignore its own programming – how are we to gain information regarding this momentous event? Outside appearances can hardly be expected to help us distinguish a rebellious robot from a lackadaisical one.

The situation gets much more complicated when we consider states of conflict.

Imagine that a robot is obliged to harm one human in order to prevent him from hurting another. The Laws are absolutely inadequate in this case. The robot should either establish an empirical hierarchy of injuries – or an empirical hierarchy of humans. Should we, as humans, rely on robots or on their manufacturers (however wise, moral and compassionate) to make this selection for us? Should we abide by their judgment which injury is the more serious and warrants an intervention?

A summary of the Asimov Laws would give us the following "truth table":

A robot must obey human commands except if:

Obeying them is likely to cause injury to a human, or
Obeying them will let a human be injured.
A robot must protect its own existence with three exceptions:

That such self-protection is injurious to a human;
That such self-protection entails inaction in the face of potential injury to a human;
That such self-protection results in robot insubordination (failing to obey human instructions).
Trying to create a truth table based on these conditions is the best way to demonstrate the problematic nature of Asimov's idealized yet highly impractical world.

Here is an exercise:

Imagine a situation (consider the example below or one you make up) and then create a truth table based on the above five conditions. In such a truth table, "T" would stand for "compliance" and "F" for non-compliance.

Example:

A radioactivity monitoring robot malfunctions. If it self-destructs, its human operator might be injured. If it does not, its malfunction will equally seriously injure a patient dependent on his performance.

One of the possible solutions is, of course, to introduce gradations, a probability calculus, or a utility calculus. As they are phrased by Asimov, the rules and conditions are of a threshold, yes or no, take it or leave it nature. But if robots were to be instructed to maximize overall utility, many borderline cases would be resolved.

Still, even the introduction of heuristics, probability, and utility does not help us resolve the dilemma in the example above. Life is about inventing new rules on the fly, as we go, and as we encounter new challenges in a kaleidoscopically metamorphosing world. Robots with rigid instruction sets are ill suited to cope with that.

Sam Vaknin is the author of Malignant Self Love - Narcissism Revisited and After the Rain - How the West Lost the East. He is a columnist for Central Europe Review, United Press International (UPI) and eBookWeb and the editor of mental health and Central East Europe categories in The Open Directory, Suite101 and searcheurope.com. Visit Sam's Web site at http://samvak.tripod.com

ThinkExist.com: Fourth of July Quotes
by: **Mark A. Lugris**

MADRID – For millions of Americans, this Fourth of July will be a time of celebration and remembrance of those passed or distanced by war. The annual commemoration of Independence Day has always been a bittersweet time for Americans who evoke the thoughts of Cordell Hull, "I am certain that, however great the hardships and the trials which loom ahead, our America will endure and the cause of human freedom will triumph."

First celebrated in 1776 and declared a legal holiday in 1941, the Fourth of July is truly the most American of all holidays. A time for families and friends to gather, the Fourth of July has become a benchmark between the blooms of spring and Indian summer.

Our national holiday is described best in the words of our most prominent leaders, past and present.

"Where liberty is, there is my country." - Benjamin Franklin

"Only our individual faith in freedom can keep us free." - Dwight D. Eisenhower

"The experience of democracy is like the experience of life itself-always changing, infinite in its variety, sometimes turbulent and all the more valuable for having been tested by adversity." - Jimmy Carter

This year as millions of Americans celebrate at backyard barbeques and on beaches across the country, we will remember the thousands who are far from home, and we will seek solace in the words of our forbearers. As Elmer Davis, the legendary American radio announcer and news commentator, stated, "This will remain the land of the free only so long as it is the home of the brave."

For more Fourth of July quotes, visit www.thinkexist.com

Mark A. Lugris is the Public Relations Director for ThinkExist.com
mark.lugris@thinkexist.com

Related eBooks:

ThinkExist.com: Fourth of July Quotes
Christ and Covenant
The Science of Robosapien
Dalton's law and diving
Reciprocal Links: The Net's Powerful Little Secret

Get more Free PDF eBooks at FreePDFeBooks.com

Related Products:

How to Gain and Retain More Customers
All Christian Writings
How to play a Guitar
English Slang Dictionary
PPC Profits

Malamaal.com: A genuine resource center for Quality Ebooks and Softwares