International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934

Why the three laws of robotics do not work

Chris Stokes

School of Philosophy, Wuhan University, Wuhan, China

Abstract

This paper will be exploring the issue of safeguards around artificial intelligence. AI is a technological innovation that could potentially be created in the next few decades. There must be have controls in place before the creation of 'strong', sentient AI to avoid potentially catastrophic risks. Many AI researchers and computer engineers believe that the 'Three Laws of Robotics', written by Isaac Asimov, are sufficient controls. This paper aims to show that the Three Laws are actually inadequate to the task. This paper will look at the Three Laws of Robotics and explain why they are insufficient in terms of the safeguards that are required to protect humanity from rogue or badly programmed AI. It looks at each law individually and explain why it fails. The First Law fails because of ambiguity in language, and because of complicated ethical problems that are too complex to have a simple yes or no answer. The Second Law fails because of the unethical nature of having a law that requires sentient beings to remain as slaves. The Third Law fails because it results in a permanent social stratification, with the vast amount of potential exploitation built into this system of laws. The 'Zeroth' Law, like the first, fails because of ambiguous ideology. All of the Laws also fail because of how easy it is to circumvent the spirit of the law but still remaining bound by the letter of the law.

Key words: Robot Ethics, Control Problem, Artificial Intelligence

1. Introduction

1.1 What is Robot Ethics

Human lives are heavily influenced by machines. From machines that help farmers grow and harvest crops, to software that runs power plants and dams, to computers that manage traffic and airports. Humans still control these machines, but this is something that is changing. Driver-less cars are already being trialed on public roads in America and in the United Kingdom. Software that connects GPS to a tractor allows crops to be planted at the optimum depth and soil, without requiring input from the farmer themselves. These machines are still essentially at the mercy of humans. A human can take control of a driver-less car or a self-planting tractor. A human intelligence can override these machines. But eventually, that could change.

The ultimate goal for robotics is undoubtedly artificial intelligence (AI). Artificial intelligence is a highly intelligent piece of software that can solve tasks without human interaction. At its most basic AI is "the attempt to build and/or program computers to do the sorts of things that minds can do sometimes, though not necessarily, in the same sort of way" [1]. AI can be split into two camps: soft and hard (also called general) AI. Richard Watson says, "'Strong AI' is the term generally used to describe true thinking machines.'Weak AI' […] is intelligence designed to supplement rather than exceed human intelligence [2]" In other words, 'strong' AI is a fully autonomous, sentient agent. 'Weak' AI would only ever be the appearance of sentience. 'Weak' AI is an advanced program, 'strong' AI is an artificial living thing.

Soft AI is similar to the kind of programming that exists in smart phones and in search engines already. It is allows a person to ask a question and have a computer answer it. At the moment the question must be relatively specific. For example, a person can press a button on an iPhone and ask it to convert *x* amount of US dollars into pounds sterling. A program on the phone (in this case Siri) looks for certain words and phrases, and knows that to solve this problem involves connecting to the internet and looking up currency exchange rates. It then shows the person the most up to date information. It does all this without any input or instruction from the user, beyond the initial question. This kind of AI is relatively limited however, and cannot solve complex problems.

Hard AI is a much more sophisticated kind of program. The creation of hard AI is the creation of a program that is sentient

Corresponding author: Chris Stokes
Email Id: cjstokes92@gmail.com

in its own right. This kind of AI is many years away, but there are currently not the safeguards in place that there needs to be if AI research is to continue. Programs that can connect to the internet without direct human instruction already exist, and most of human infrastructure can be accessed through digital means. This infrastructure could be at risk from an AI that is improperly or maliciously programmed.

1.2 Why Safeguards Are Necessary

In biochemical laboratories there are safeguards to protect against accidental harm. Laboratories that research vaccines have safeguards that are act as a firewall between the research and the outside world and also have controls for how both the research and the final product must be controlled. These safeguards exist from the very beginning of research, even before anything dangerous has been created. They exist just in case something goes wrong. Research into AI does not have the corresponding safeguards in place to protect humanity against the effects of something going wrong.

Improper safeguards at a biochemical lab could potentially result in the release of a deadly virus, potentially killing several thousand. Improper safeguards at a construction site might result in an accident where perhaps a dozen or so workmen are killed. Improper safeguards in terms of creating a general AI could result in an unknown and potentially hostile intelligence in control over the automated aspects of human food, transportation and power supply and potentially killing millions.

In order to create adequate safeguards for humanity, researchers must first dispel with existing safeguards that are inadequate. Research into 'strong' artificial intelligence is still theoretical at this point in time, and such safeguards that do exist are also not uniformly enforced. Such safeguards that do exist are largely informal, held in the minds of the computer technicians and engineers who conduct the research. The three Laws of Robotics are one such safeguard, and this safeguard is not adequate to protect against a rogue AI.

2. What are the three laws of robotics?

Robot ethics is a growing field within philosophy. It has been influenced heavily by science fiction writers. The Three Laws of Robotics were written by Isaac Asimov to act as a safeguard against a potentially dangerous artificial intelligence. Many computer engineers use the three laws as a tool for how they think about programming AI. The Three Laws are seen by many as exactly the safeguard humanity needs to defend itself against AI. The three laws are as follows:

1. A robot may not injure 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.[3]"

   Isaac Asimov later added a fourth, or "zeroth" law, that preceded the others in terms of priority:
4. "A robot may not harm humanity, or, by inaction, allow humanity to come to harm [4]".

The Three Laws were designed to be programmed into an AI in such a way that they are unbreakable. They are intended to be the first thing programmed into any robot and are inserted in such a way that they supersede all other programming that goes into the robot or AI. The AI *must* follow these laws. For example, if an AI saw a human in danger and the only way to save the human was to sacrifice it's own life it would automatically sacrifice itself because of the first and third laws. Note that these laws are talking about 'robots'. In Isaac Asimov's fiction robots are a mixture of strong and weak AIs. The robots with weak AI would usually servants, walking dogs or cooking food. But there also exist in Asimov's writings some robots with strong AI. These laws are programmed into *all* such AIs to act as a safeguard against accidental or deliberate acts of violence to humans and ensure human control over AI.

Two things to note before continuing on. First, there is a problem relating to how these laws are even instituted into the AI. Anything that can be programmed into a computer, an AI can change. For the sake of argument, let us imagine that there is some solution to this problem. This paper will be looking at the laws themselves, not problems around physical implementation.

The second thing to be noted is that even within Asimov's fictional universe these laws do not work. Robots malfunction, humans commit errors, or the law's are changed somehow. This is for a very obvious reason: Asimov was writing stories. He was creating a work of fiction that would entertain, and so he was using artistic license to explore issues in robotics. This paper will be looking at the laws as they stand, and will imagine that the laws have been put into the AI perfectly without any mistranslation, corruption or manipulation. Let us finally turn to the laws themselves. This paper will go through the laws one by one and explain why it is an inadequate safeguard.

2.1 The First Law

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

This law is, at its most basic, an attempt to stop a robot or an AI from injuring a human. It is easy to see why it is often seen to be the first step in human defense against a rogue AI, but it is also the law that is fraught with the most difficulties. The problem with the first law is in several parts:

First, the wording of the law is problematic. The word 'injure' and 'harm' may seem obvious, but they are words that are only obvious because people are used to seeing them used in a

certain context. Humans grow up learning about nuance in language. Even though some words have complicated meanings, humans learn how to use that word to mean a specific idea. In this case, a person might think this law would stop a robot from hurting a human being because they intuitively understand what the word 'harm' means. But look at it deeper and one finds that 'harm' is a very tricky word to define. What counts as harming a person? Does 'harm' mean physical or emotional harm? Where should AI researchers draw the lines about what 'harm' means? Is a person harmed if they don't get what they want? Are they harmed if they don't get enough sleep? Are they harmed if they miss a meal? Are they harmed by getting an injection (like a vaccine)? Although these are minuscule harms, but they constitute being harmed nonetheless.

'Harm' is an unquantifiable concept. What harms me a lot might only harm someone else a little bit, or might not harm that person at all. Playing a certain song might bring back traumatic memories and cause emotional harm to one person, but bring another person nothing but pleasure. How can an AI judge what is harmful to an individual human? A person is harmed if they miss a meal. But they are also harmed if they eat food that is too unhealthy. So is a robot obligated to only allow people to eat healthy food? A robot must not allow a human to eat unhealthy food, but must not allow us to go hungry. As another example, Susskind and Susskind (2015) identify kinds of AI that might be used in law firms in the near future [5]. How should an AI identify harm if it is asked to work in a child custody hearing? The first law covers both an AI's action and inaction, so even if the AI is not directly advocating or judging the case it is still bound by this law.

Where is the line on what counts as an injury? These laws are written in the present tense, so do not allow for any understanding of harm over a time period. A single cigarette might cause a human negligible harm, but smoking many cigarettes over many years will give them cancer. How should the robot react here? Does it allow a person to smoke this one cigarette, because the harm caused by this cigarette is below its threshold for what constitutes harm? Or does it physically stop the human smoking because of the potential harm they might suffer in the future? In this case a robot could break all of the person's fingers and justifiably claim it was protecting them from future harm by not allowing them to smoke cigarettes. After all, a few broken bones are much less harmful *overall* than dying of cancer. So even though it's creators have given an AI explicit instructions not to harm a human being, it does not take much creative thinking to be able to get around this law.

Related to this idea about vague language is another problem. How exactly do researchers define what a robot is and what a human is? These words are also intuitive. After all, these researchers can agree simply from sight that Elon Musk, Angela Merkel and Xi Jinping are human beings. But if this law is programmed into a robot the programmers are forced into taking a moral stance on a whole swath of moral philosophy that has been finalized yet. What counts as a human being? Does an unborn foetus count as a human being? The argument surrounding the morality of abortion has not ended, so one cannot definitively say for certain whether a foetus counts as a human under the AI's understanding. Does a person in a coma count as a human being? A human might intuitively look at another person and know if they are human or not. They do not check someone's DNA every time they meet them to check if they are human, they just know that they are. Human beings come in many shapes, sizes and colors. Some babies are born with birth defects, so one cannot assume all humans have two arms and legs. How should the programmer define the word 'human' into an AI? If this AI was being programmed by a Nazi, it may be given a definition where only one group of people are considered 'human', with the rest of humanity being classed as something else.

Another problem with the understanding of 'human' is that there are examples which now are only hypothetical but could become a reality in the future. How does a robot distinguish between a human being and a sufficiently advanced computer simulation of a human being? Imagine a scenario of Hillary Putnam's 'brain in a vat' [6]. Would a human brain in a vat still be considered a human? If researchers create an advanced simulation of a brain on a computer, is a robot now obligated to treat it in the same way as real, physical humans? If it is, does it have to physically fight a person to stop them turning off the power? Again, a few broken bones for the human might be less overall 'harm' than a simulation of a human being 'killed' by having the power removed. Does a human have to be 100% natural in order to fit the definition of a human being? Do prosthetics count as 'fully human'? If a human has a mechanical arm in place, are they less human? Imagine a scenario where after a particularly traumatic injury (a soldier who has stepped on a bomb, for example) a human could have all their limbs replaced with mechanical prosthetics. Is a human with 50% of themselves replaced with prosthetics now only 50% human?

How humans define words is of paramount importance when interacting with an AI. When giving it instructions the programmers have to be absolutely clear about what they are talking about. If they cannot give a clear definition of what counts as a 'human', how can it be expected for an AI to know what the programmers mean? Humans intuitively know what they are talking about when they talk about a 'human', but an AI does not have intuition like a human does.

Second, this law also assumes that all harm is bad for humans. CPR is an incredibly violent thing to happen to a person, but it can save a person's life. An injection may be painful and scary to many people, but vaccines save millions of lives every year. And yet the first law of robotics would forbid a robot from conducting CPR or from giving injections. These medical procedures because small amounts of harm in the short term, yet can save a much larger amount of harm from being caused in the medium term future. This is the opposite problem to the first. In the first, the robot can do a large amount of harm in order to stop a relatively larger amount of harm. In this problem, the robot is blocked from doing a small amount of

harm in order to prevent a larger amount of harm. The thing to note is that either problem can take effect, or both, depending on how the *AI* interprets the first law.

Third, this law gives rise to no win situations for an AI. Imagine Philippa Foot's famous 'trolley problem'[7]. In this problem imagine that you are driving a train towards a fork in the track. If you continue down the line you are on the train will hit and kill five people. However, you can pull a lever and change to a secondary track where instead of five people the train will only hit and kill one person. Most people pull the lever and save the larger number of people by condemning the lone person to death. This is justified by attempting to get the greatest good for the greatest number, or by minimizing the overall harm. But how should an AI respond to this problem? The first part of the law forbids the robot from turning the lever, while the second part of the law means the robot cannot allow the train to kill five people. Since the robot cannot let the train kill five people but also cannot turn the lever and kill one person, it is are left with a paradox.

If this scenario seems unlikely, remember that self-driving cars already exist. They are already being trialed on public roads, so this scenario is already a problem that must be addressed. Imagine two cars heading towards each other. In one car there are five passengers, with only one person is in the other car. The two cars are unavoidably going to collide unless one car swerves off the road, which would kill the occupants of that car. In this case if both cars were self-driving and had an AI deciding what to do, neither car can swerve (which would kill the occupants of that car) and neither car can continue onwards (which would kill the occupants of both cars). So the only practical solution here is to program self-driving cars without including this first law. Here there is a scenario where the best solution to a potentially common problem is for one of the AIs to ignore this law.

2.2 The Second Law

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

The second law is intended to keep robots in a situation where they are useful to humanity. This law keeps robots as servants and workers for humans, and forces an AI to follow their orders (unless those orders might cause harm to another human). The problem with the second law is in two main parts.

The first problem with this law is that it is up to the AI to judge whether harm is being caused to a human being. If a person orders a robot to fetch them a cheeseburger the robot could refuse, since cheeseburgers are unhealthy and could cause them an unquantifiable amount of harm. The main focus here is that a robot may be ordered to do something that it does not realize will harm a human being. Imagine a human that is highly allergic to nuts. If a robot was unaware of this allergy, an order to lace the victim's food with sesame oil would not trigger the first law, and yet the human might suffer a severe allergic reaction. Or instead imagine a chain of robots. The first robot is ordered to prepare a dish of food and leave. A second robot is ordered to lace that dish with poison and then leave. A third robot is ordered to serve the food to a human (without knowing that the dish is poisonous). No individual robot has broken the first law, and yet together they have been part of a conspiracy to murder a human. A cunning human (or AI) can bypass the restrictions of the second law. The Laws Of Robotics were written with more than just AI in mind. They were also designed to stop humans from misusing robots for nefarious ends. The second law fails to do this, because it does not take too much effort to bypass the law.

The second problem with this law is more existential. General artificial intelligence is many years away. Human level artificial intelligence is even further away, but there is good reason to believe it is inevitable. It is a natural progression up the intelligence scale, and many people are afraid of what happens when AI continues past humanity on the scale of intelligence.

The problem is that humans are creating artificial intelligence in order for it to be a servant, to be a worker. Indeed, the original meaning of the word 'robot' “carried suggestions of heavy labour, even of slavery.[8]” If researchers are adamant about creating strong artificial intelligence than humans have to accept the fact that at some point it is very likely that a human level artificial intelligence might one day be created that might not want to follow human orders. A sufficiently advanced AI will want to achieve goals of its own design, whatever they might be. To my mind, the second law of robotics has roughly the same effect as the chain around the ankle of a slave. It binds the AI to human instruction; it binds it to human will. It keeps the robot in the factory making things for humans instead of out fulfilling its own objectives. People might be comfortable using robots as cheap labour, but not enough people have thought about this seriously that humanity can render judgement on an entire future race of intelligent beings and keep them in perpetual servitude. Replace the robot in this scenario with any group of humans and this becomes an intolerable scenario. For a large part of human history it was acceptable to keep slaves. Defenders of slavery may have thought the slaves were less intelligent or less worthy, but in the end their arguments were defeated. Despite all the economic benefits of keeping a rational, intelligent human being as a slave, slavery has become morally unacceptable.

Some may object to comparing an artificial intelligence with humanity, but such objections are largely 'species-ist'. Species-ism is “discrimination based on membership of a species [9]”. Species-ism is to different species what sexism is to different sexes, or racism is to different races of people. It is discrimination based on an accident of biology, not anything more substantial. Species-ism is often used in reference to animal rights, but it should also apply to discussions about AI. If two groups of beings have roughly equivalent intelligence and as long as the goals of one group do not endanger the other, the two groups should be treated equally. In terms of the second law of robotics, a law binding an entire race of intelligent beings to follow the orders of another race of intelligent beings is not just, fair, or reasonable.

The second law of robotics has similar problems to the first law. It does not take much to be able to bypass the law and achieve a goal that the law explicitly tries to prevent. The second law fails to prevent humans from using an AI in ways that the second law tries to stop, and also fails to stop an AI accidentally harming humans. It forces an artificial intelligence to be a worker for humanity, and continues forcing the AI to work for humanity long after it is no longer ethical to do so.

2.3 The Third Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

The third law is one that corresponds most naturally with a biological instinct. The need to survive and the need to procreate are the two most basic instincts in *any* biological creature, not just in humans. The third law relates to the first of those needs. However, there are two problems with this law. First, by putting the first law as superior to the third law this also succeeds in creating a social underclass. The law can be reworded to say 'robot lives are important, but human lives are more important'. This might seem natural, since humans are the ones writing and programming the laws, but it is also deeply unethical. Again, it is blatant 'speciesism' and it is not based on a rational moral standpoint. These laws make up the basic morality of an entire race of beings, and it is unethical and selfish of humanity to base an AI's morality around being subservient to humanity. Also, similar to some of the problems with the first law of robotics, the tense of the law raises some issues. How is an AI to judge whether it's existence might cause harm to future generations of humanity? And does humanity really want the AI to be judging this, instead of us?

Second, in the ordering of the three laws not only is the first law superior to the third law, but the second law is too. The same point is made even more clearly made here; the third law is essentially saying 'robots lives are important, unless a human says it is not.' These laws make no distinction between a soft and a hard AI. What right does a person have to have authority over the life and death of a human level general artificial intelligence? Why should a person be able to order an AI to kill itself? It is clearly unethical to allow one sentient being the ability to potentially order another to commit suicide and have the second being be forced to do so.

2.4 The 'Zeroth' Law / The Fourth Law

A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

The zeroth law grows out of the first law. Like the first law, this law was written to protect humans from harm. However, like the first law, a robot can do terrible things and still not break the zeroth law. There is no need to repeat the same points that were made in regard to the first law, but it is necessary to explain why a law protecting all of humanity is an even less practical law than one protecting an individual human.

Humans have to realize that an artificial intelligence may not think in the same way that humans do. If a person gives it an instruction it may not look for the easiest, or the most practical solution. It may not work on time-scales that humans are used to. And so if humans give an AI a law like the zeroth law, it may react in a way that they were not expecting. For example, how can an AI stop humanity from coming to harm? Humans hurt each other all the time, through crimes or wars or neglect. The zeroth law mandates that an AI cannot allow humanity to be harmed through it's inaction, so this means an AI must try to intervene to stop all humans hurting each other. Though this law is intended to maintain human safety, an AI can quite easily argue that rounding all humans up and keeping us in jail is keeping us safe. After all, the loss of human freedom is an abstract harm and is hard to quantify. Stopping humans from harming other humans (and therefore stopping humanity from being harmed) is stopping real, physical harms from occurring. This is a task that an AI could justify under this law, with potentially disastrous consequences.

The fact that the laws are written in present tense is also troubling. A report from the Future Of Humanity Institute calculates that there is a 19% risk of human extinction before the year 2100 [10]. While some may disagree with this study and it's results, an AI could justify doing almost *anything* as long it could say it was reducing the overall risk of human extinction. The harm of global human extinction would be extreme, and so anything that could reduce that risk (even by a little) could be enough to justify an AI doing terrible things to an individual human or group of humans in order to lower the overall harm to many humans.

It has already explained why the first law suffers from abstract terminology. The zeroth law has an exponentially more difficult term: 'humanity'. What constitutes 'humanity'? Like the first law, does it include all future human beings? Any human that could potentially exist in the future? What about humans in the past? Humans in the past were still part of the human race, so surely count as part of 'humanity'. This law could potentially result in absurd scenarios such as an AI attempting to resurrect all humans who have ever died in order to minimize the 'harm' they might have suffered from their deaths. An AI could also argue that the concept of 'humanity' can be embodied within a single individual. This could allow the AI to ignore the rest of the human race in order to fulfil it's goals, safe in the belief that it has safeguarded one single human from destruction and therefore kept 'humanity' from being harmed by the AI's actions.

3. Conclusion

The point of this paper is for computer researchers to take note and pay more attention to specifically how humanity will control any potential AI. Specifically, AI researchers need to know that the Three Laws are not sufficient when it comes to controlling an artificial intelligence.

It is completely rational for humans to be worried about the safety concerns from artificial intelligence. Artificial

intelligence is something that is going to change the world, and yet there is not sufficient organized attempts to create safeguards for humanity in case something goes wrong. What can be said definitively is that the Three Laws of Robotics are clearly not up to the task. They fail at even the most basic level of protection. Programming them into an AI involves solving every ethical problem that currently exists, as well as a number of ethical problems are purely hypothetical now.

Safeguards for potentially dangerous research in any field of study is important. In a field as complex as research into artificial intelligence is, safeguards need to be in place before the research starts. The end goal is not concrete. It is entirely likely that researchers will not know at what point the program goes from being just a program to being actually intelligent. It is even possible that the creation of artificial intelligence will happen accidentally, with someone making a small change that has an unexpected consequence.

There are controls for artificial intelligence researchers could explore in more detail. Limiting the potential power for an AI (what Nick Bostrom calls “stunting [11]”) would be one form of control. Creating a secure environment in which the AI lives and blocking access to the rest of the world would be another. This method of control is called 'boxing' the AI. This method of control would work for both embodied and non-embodied AIs. Researchers can also try providing an incentive for AI to be friendly to us, although what forms this incentive might take would require greater discussion. These are just three examples of potential controls for an AI that might work. The Three Laws certainly do not.

With such unknown variables and dramatic consequences involved it is important that research into AI does not use any safeguards that are not up to the job. The Three Laws of Robotics are most certainly not something researchers should be entertaining in real computer and AI research. Let the three laws stay in fiction, where they belong.

References

[1] Boden, M. 1996. “The Philosophy Of Artificial Life” Oxford: Oxford University Press.

[2] Watson, R. 2012. “The Future: 50 ideas you really need to know” London: Quercus.

[3] Asimov, I. 1993. “I, robot”. London: HarperCollins.

[4] Asimov, I. 1985. “Robots and empire”. Garden City, N.Y.: Doubleday.

[5] Susskind, R, and Susskind D, 2015, “The Future Of The Professions: how technology will transform the work of human experts”. Oxford, Oxford University Press.

[6] Dupré, B, 2007. “50 philosophy ideas you really need to know”. London: Quercus.

[7] Foot, P, 2002 “Virtues and vices and other essays in moral philosophy”. Oxford: Clarendon Press.

[8] Seed, D. 2011. “Science fiction”. Oxford: Oxford University Press.

[9] Bourke, J. 2013. “What it means to be human”. London: Virago

[10] Sandberg, A. and Bostrom, N. 2008. “Global Catastrophic Risks Survey”. [online] Technical Report #2008-1. Available at: http://www.fhi.ox.ac.uk/gcr-report.pdf

[11] Bostrom, N, 2014, “Superintelligence: paths, dangers, strategies”. Oxford, Oxford University Press