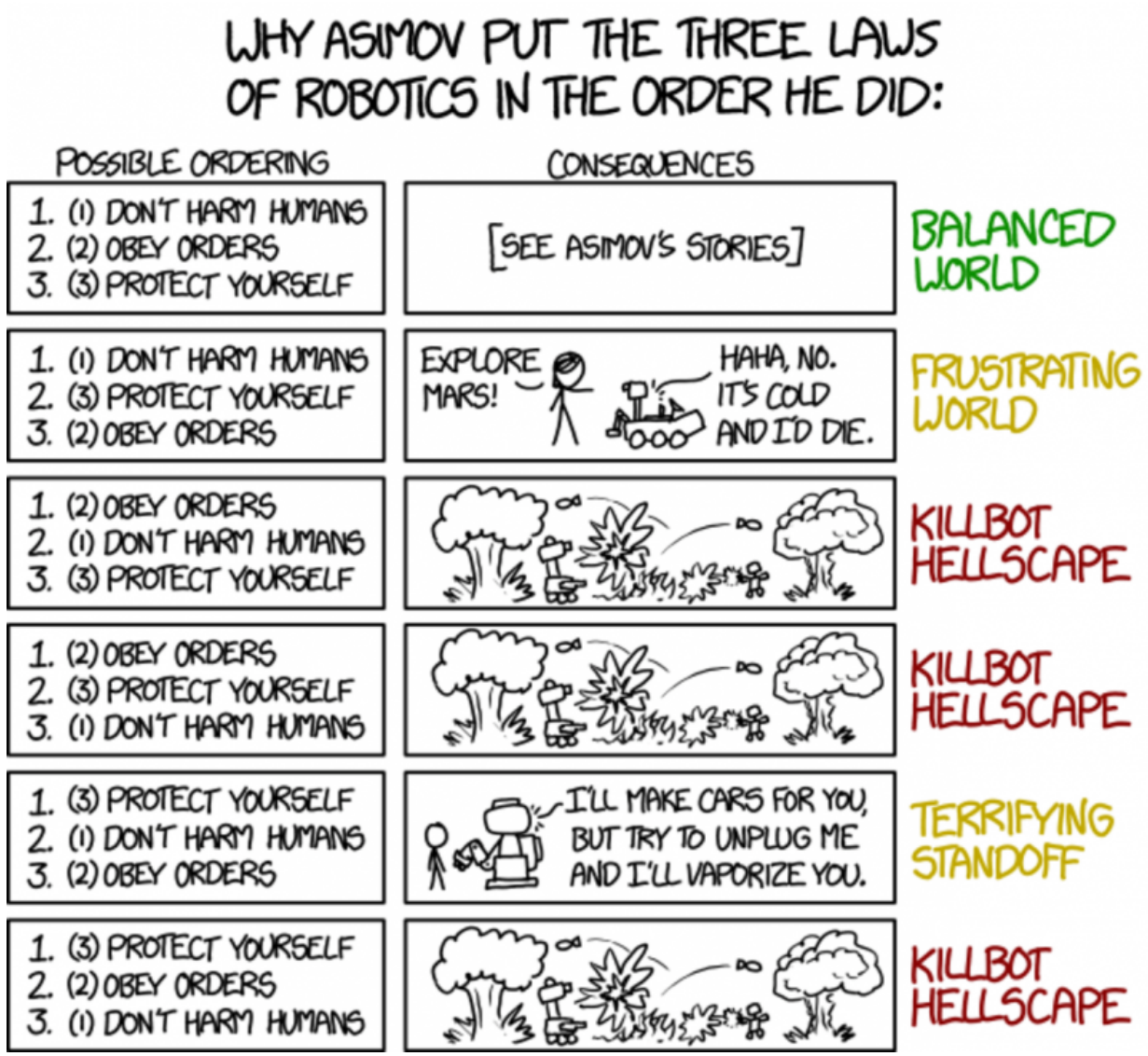
Beyond Asimov: how to plan for ethical robots

June 2 2016, by Benjamin Kuipers, University Of Michigan



Asimov's laws are in a particular order, for good reason. Credit: Randall Munroe/xkcd, CC BY-NC

As robots become integrated into society more widely, we need to be sure they'll behave well among us. In 1942, science fiction writer Isaac Asimov attempted to lay out a philosophical and moral framework for ensuring robots serve humanity, and guarding against their becoming destructive overlords. This effort resulted in what became known as Asimov's 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.

Today, more than 70 years after Asimov's first attempt, we have much more experience with robots, including having them drive us around, at least under good conditions. We are approaching the time when robots in our daily lives will be making decisions about how to act. Are Asimov's Three Laws good enough to guide robot behavior in our society, or should we find ways to improve on them?

Asimov knew they weren't perfect

Asimov's "I, Robot" stories explore a number of unintended consequences and downright failures of the Three Laws. In these early stories, the Three Laws are treated as forces with varying strengths, which can have unintended equilibrium behaviors, as in the stories "Runaround" and "Catch that Rabbit," requiring human ingenuity to resolve. In the story "Liar!," a telepathic robot, motivated by the First Law, tells humans what they want to hear, failing to foresee the greater harm that will result when the truth comes out. The robopsychologist Susan Calvin forces it to confront this dilemma, destroying its positronic

brain.

In "Escape!," Susan Calvin depresses the strength of the First Law enough to allow a super-intelligent robot to design a faster-than-light interstellar transportation method, even though it causes the deaths (but only temporarily!) of human pilots. In "The Evitable Conflict," the machines that control the world's economy interpret the First Law as protecting all humanity, not just individual human beings. This foreshadows Asimov's later introduction of the "Zeroth Law" that can supersede the original three, potentially allowing a robot to harm a human being for humanity's greater good.

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

Robots without ethics

It is reasonable to fear that, without ethical constraints, robots (or other artificial intelligences) could do great harm, perhaps to the entire human race, even by simply following their human-given instructions.

The 1991 movie "Terminator 2: Judgment Day" begins with a well-known science fiction scenario: an AI system called Skynet starts a nuclear war and almost destroys the human race. Deploying Skynet was a rational decision (it had a "perfect operational record"). Skynet "begins to learn at a geometric rate," scaring its creators, who try to shut it down. Skynet fights back (as a critical defense system, it was undoubtedly programmed to defend itself). Skynet finds an unexpected solution to its problem (through creative problem solving, unconstrained by common sense or morality).

Less apocalyptic real-world examples of out-of-control AI have actually taken place. High-speed automated trading systems have responded to

unusual conditions in the stock market, creating a positive feedback cycle resulting in a "flash crash." Fortunately, only billions of dollars were lost, rather than billions of lives, but the computer systems involved have little or no understanding of the difference.

Toward defining robot ethics

While no simple fixed set of mechanical rules will ensure ethical behavior, we can make some observations about properties that a moral and ethical system should have in order to allow autonomous agents (people, robots or whatever) to live well together. Many of these elements are already expected of human beings.

These properties are inspired by a number of sources includingthe Engineering and Physical Sciences Research Council (EPSRC) Principles of Robotics andrecent work on the cognitive science of morality and ethics focused onneuroscience,social psychology, developmental psychology andphilosophy.

The EPSRC takes the position that robots are simply tools, for which humans must take responsibility. At the extreme other end of the spectrum is the concern that super-intelligent, super-powerful robots could suddenly emerge and control the destiny of the human race, for better or for worse. The following list defines a middle ground, describing how future intelligent robots should learn, like children do, how to behave according to the standards of our society.

- If robots (and other AIs) increasingly participate in our society, then they will need to follow moral and ethical rules much as peopledo. Some rules are embodied in laws against killing, stealing, lying and driving on the wrong side of the street. Others are less formal but nonetheless important, like being helpful and cooperative when the opportunity arises.

- Some situations require a quick moral judgment and response – for example, a child running into traffic or the opportunity to pocket a dropped wallet. Simple rules can provide automatic real-time response, when there is no time for deliberation and a cost-benefit analysis. (Someday, robots may reach human-level intelligence while operating far faster than human thought, allowing careful deliberation in milliseconds, but that day has not yet arrived, and it may be far in the future.)
- A quick response may not always be the right one, which may be recognized after feedback from others or careful personal reflection. Therefore, the agent must be able to learn from experience including feedback and deliberation, resulting in new and improved rules.
- To benefit from feedback from others in society, the robot must be able to explain and justify its decisions about ethical actions, and to understand explanations and critiques from others.
- Given that an artificial intelligence learns from its mistakes, we must be very cautious about how much power we give it. We humans must ensure that it has experienced a sufficient range of situations and has satisfied us with its responses, earning our trust. The critical mistake humans made with Skynet in "Terminator 2" was handing over control of the nuclear arsenal.
- Trust, and trustworthiness, must be earned by the robot. Trust is earned slowly, through extensive experience, but can be lost quickly, through a single bad decision.
- As with a human, any time a robot acts, the selection of that action in that situation sends a signal to the rest of society about how that agent makes decisions, and therefore how trustworthy it is.
- A robot mind is software, which can be backed up, restored if the original is damaged or destroyed, or duplicated in another body. If robots of a certain kind are exact duplicates of each other, then trust may not need to be earned individually. Trust

earned (or lost) by one robot could be shared by other robots of the same kind.
- Behaving morally and well toward others is not the same as taking moral responsibility. Only competent adult humans can take full responsibility for their actions, but we expect children, animals, corporations, and robots to behave well to the best of their abilities.

Human morality and ethics are learned by children over years, but the nature of morality and ethics itself varies with the society and evolves over decades and centuries. No simple fixed set of moral rules, whether Asimov's Three Laws or the Ten Commandments, can be adequate guidance for humans or robots in our complex society and world. Through observations like the ones above, we are beginning to understand the complex feedback-driven learning process that leads to morality.

This article was originally published on The Conversation. *Read the* original article.

Source: The Conversation