

Robot Trickery

Carson Reynolds and Masatoshi Ishikawa

University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

{carson,ishikawa}@k2.t.u-tokyo.ac.jp

Introduction

As robots gain abilities that exceed those of humans a particularly interesting possibility arises. Namely, robots can trick or deceive humans without detection. Such deceptive uses of robotics lead to interesting questions concerning fault, responsibility, and agency. In this article, we approach the problem of robot ethics by speculating some novel ways in which robots might be bad.

The 3-Card Monte Robot

Three card monte is a hustle, an unwinnable game. At the outset of the game a player is shown the 3 cards, one of which is the Queen. A successful hustler uses extreme speed to shuffle the cards and confuse the player as to the location of the Queen. In the case that the player somehow managed to correctly track the cards, often accomplices step in so that the bet is refused (“one bet at a time”) or the dealer may use sleight of hand to introduce a new card. In such a way, the hustler is assured that he can win in every case (McLeod, 2001).

Imagine that a robot were made to act as dealer for 3-Card Monte. Further imagine that this robot were capable of faster-than-the-eye movement (Kaneko et al., 2003). The robot could ensure then that the player could not possibly follow the Queen card. A particularly dexterous variant might even be able to employ sleight of hand. More fantastically, a robot could move so fast that it could retrieve the wallet or purse of the player and remove money in a barely perceptible blur of motion.

The question we wish to pose to readers is: would such a robot be bad? One might argue that it is functionally equivalent to a slot machine that can adjust odds dynamically. Or one might argue that the designer’s intent in making such a robot is to deceive people. But is the robot at fault for its ability to deceive?

Robotics and Accountability

Before trying to fully answer such weighty questions, we would first like to discuss some existing literature about ethics as it applies to computer and robotic systems. Indeed a great deal of written and acted fiction has tried to grapple with the collision of robotics with human ethical

concerns: Capek’s RUR (Rossum’s Universal Robots), Asimov’s I, Robot, and Clarke’s 2001: A Space Odyssey for instance. However, these fictive accounts do not have the present perspective which is tempered with actual examples of harm induced by semi-autonomous systems.

A recent article in The Economist reported on the following unfortunately relevant historical event:

In 1981 Kenji Urada, a 37-year-old Japanese factory worker, climbed over a safety fence at a Kawasaki plant to carry out some maintenance work on a robot. In his haste, he failed to switch the robot off properly. Unable to sense him, the robot’s powerful hydraulic arm kept on working and accidentally pushed the engineer into a grinding machine. His death made Urada the first recorded victim to die at the hands of a robot (Anonymous, 2006).

Another incident that is worth revisiting is that of the Therac-25 chemotherapy machine. An investigation by Leveson and Turner reports “Between June 1985 and January 1987, six known accidents involved massive overdoses by the Therac-25 – with resultant deaths and serious injuries.” A software error resulted in the machine administering overdoses.

In trying to understand who or what is accountable in the Therac-25 incidents, one study interviewed 29 participants (Friedman, 1995). The participants were computer science majors who were asked about scenarios, including one involving a “a computer error” which “over-radiates a cancer patient.” Friedman found that 83% “attributed aspects of agency” to the computers. While 21% held the computers “morally responsible.”

To build upon this work, it would be enlightening to conduct similarly structured interviews that discuss scenarios involving robotics. For instance, if a faulty decision due to misprogramming causes the death of a person by a robot, did the robot make the decision or did the programmer? More importantly, is the robot morally responsible for the decision?

Designer or Robot

The most likely candidates for the “morally responsible” entity would be the robot itself or the designer. As an aside, one

might argue that the moral responsibility is a socially constructed phenomena and thus really not a property of robot or the designer, but of the society as a whole. But for the present purposes, let us limit attention to the prototypical robot and the prototypical designer. Specifically, we wish to enumerate a few different cases involving intent.

The malevolent designer

Suppose that the designer of the robot is malevolent. This designer wishes to make a robot which will do harm and trick people. The goal of the designer might be further his or her own interests using a robot in an anti-social manner. For instance, the malevolent designer may wish to be wealthy and thus employ many three-card-monte robots to trick individuals into giving the designer money.

The oblivious designer

In contrast we might have the oblivious designer. This is an individual who does not view his or her work in terms of moral consequences. Perhaps this individual feels that he or she is only designing widgets to a particular specification. This designer is not concerned with eventual deceptive or undeceptive usage.

The benign designer

To round out the spectrum of prototypical designers we might have a benign designer. This designer wishes that his or her robot will not do harm. This designer might further be concerned about designs being used in ways contrary to his or her intentions.

The malevolent robot

Now let's consider a similar spectrum of robots. A malevolent robot is one which (irrespective of the designer's intent) is harmful, perhaps in a deceptive manner. Such a robot might make use of its super-human capabilities (e.g. moving faster than the eye can see) to manipulate individuals.

The oblivious robot

The oblivious robot does not conform to its designer's moral intent. Either a malevolent or a benign designer could fail in attempts to create a robot which mirrors his or her moral stance.

The benign robot

The benign robot is one which does not cause harm. This might be due to a competent benign designer or an incompetent malevolent designer. Perhaps an example of such a robot might be a robotic surgery system that makes use of super-human capabilities to perform live-saving operations.

The Permutations

In examining the different permutations of a intent of designer and robot we arrive at some possible ways to frame our question about whether the deceptive robot is bad or faulty. Intuitively, if a benign designer makes a malevolent robot then the robot itself is faulty. The robot is bad because it not only doesn't behave as the designer expects, but does

harm in the process. We may similarly speculate that in the case of the malevolent designer and benign robot that the designer is bad because regardless of the outcome harm was intended on other individuals.

But this begs a critical pair of questions: "Why would robots be malevolent on their own and why would designers want to produce malevolent robots, which may harm humans?" In order for robots to be malevolent on their own it seems that they would have to have some self-serving goal (perhaps for batteries to power themselves) that would motivate malevolent behavior (robbing an electronics store). It is not as hard to think of reasons for the creation of malevolent robots by humans; one need only examine recent research concerning the use of armed autonomous robotics in warfare (Foster-Miller, 2005).

Still more interesting and complex problems arise when the possibility of the robot performing moral decision making is introduced (Gips, 1995). One might ask whether free will is required for a robot to be benign, or if a functional version of Searle's Chinese-room would suffice.

In this article we have briefly discussed the idea of robots tricking individuals through the use of capabilities that exceed those of humans. Following on this we have speculated on some varieties of intents and ethical implications. Ultimately, such speculation should be grounded in empirical work and so this paper proposes conducting interviews about such scenarios as a next step.

References

- Anonymous (2006). Trust me, im a robot. *The Economist*. http://www.economist.com/displaystory.cfm?story_id=7001829.
- Foster-Miller (2005). Army technology - robot with gun mounted. http://www.army-technology.com/contractors/civil/foster_miller/foster_miller5.html.
- Friedman, B. (1995). "it's the computer's fault": reasoning about computers as moral agents. In *CHI '95: Conference companion on human factors in computing systems*, pages 226–227, New York, NY, USA. ACM Press.
- Gips, J. (1995). Towards the ethical robot. In *Android Epistemology*, pages 243–252. MIT Press, Cambridge, MA, USA.
- Kaneko, M., Higashimori, M., Takenaka, R., Namiki, A., and Ishikawa, M. (2003). The 100 g capturing robot - too fast to see. *Mechatronics, IEEE/ASME Transactions on*, 8(1):37–44.
- McLeod, J. (2001). Rules of card games: Three card monte. <http://www.pagat.com/misc/monte.html>.