Ethics and AI

Christoph Schulze

Introduction

The ethical implications of Artificial Intelligence (AI) have been addressed in science fiction for decades. The creativity and freedom of fictional authors have allowed them to foreshadow a lot of the ethical questions that we are facing in today's computerized world. Mainly to what degree and how AI and artificial machines should participate in our society and how we should treat them in return. The interdisciplinary fields of Machine Ethics and Roboethics try to address these questions from a scientific viewpoint now.

Although those two fields by themselves are rather new the argument in the scientific community is much older. Joseph Weizenbaum, creator of the ELIZA program[1] was an advocate for the restrictive usage of AI. He claimed that certain tasks, like nursing or passing judgment, should never be done by AI's. He argued that those professions require compassion and intuition, qualities that he believed cannot be achieved by machines[2].

On the other side there are arguments to use artificial beings and AI to improve the quality of our lives. But that is not the only argument for the "pro AI" stance. Japanese culture for example has a very different and much more optimistic view about artificial life than we have in our Western culture [3][4] and therefore some of the ramifications of artificial life are viewed very differently there.

The remainder of this report will address the following topics. Joseph Weizenbaums critique of AI as an example of an argument from the "Anti AI group" as well as counterarguments to Weizenbaums view. Furthermore we will briefly discuss how ethical views and implications on AI differ between Japanese and Western culture, showing that there are different viewpoints to consider when talking about ethics and AI. The last part of the report describes the fields of Machine Ethics, which tries to explicitly embed ethics into machines, and Robethics which is concerned with the ethical treatment of artificial beings by humans.

Weizenbaums argument against Artificial Intelligence

Joseph Weizenbaum was not always distrustful of computers but one experience in his life changed his view completely. In 1966 he developed ELIZA a computer program that through very clever natural language processing could give people the impression that they are talking to a real person, for example a psychiatrist that understands their problems. He was scared by the emotional involvement that some people put into the communication with ELIZA and by researchers who actively tried to use it for psycho analysis since he knew it had no understanding of the problem of psycho analysis at all. [5]

In 1976 he published the Book Computer Power and Human Reason [2] in which he argues that machines should never be allowed to do tasks for which they lack certain qualities that he believes are essential and only humans can have (E.g compassion, intuition, creativity). He claims that there is more to the human mind than just the brain and that a simple replication of the brain will never achieve realistic human behavior.

In his view the whole history of a human being makes up its mind and that the body also plays an important role in the development of the mind[1]. Weizenbaum sees computer programs as useful tools for humans to work with but believes they should only be used if we understand them and their consequences completely.

Pamela McCorduck says[2] "I'd rather take my chances with an impartial computer", which points out one of the problems in Weizenbaums reasoning about why computers should not be judges. A computer if programmed correctly would be impartial for example towards gender or minority issues. Humans might be biased. So even if a computer based judge might be imperfect the human judge is potentially flawed as well.

John McCarthy refuses the claims of the book outright [3] and is attacking it on the grounds that Weizenbaums aversion of AI boarders on the extreme. And that the book is too moralizing and vague and even goes so far as describing this line of thinking as dangerous to science..

Joshua Lederberg pointed out[4] that he doesn't believe it is possible to only create programs whose ramifications we completely understand, like Weizenbaum requests, but that we should make sure that they behave properly by verifying their behavior. Along this line of thought Ben Schneiderman presented an idea [6] how to hand over responsibility to autonomous systems. He argues that they should be closely monitored and gives several examples on how to do so (monitoring, control and logging tools). He proposes a step wise approach where humans initially monitor the system heavily and as they gain trust in the system the monitoring can be decreased.

Perception of Artificial Intelligence in other cultures

A predominant view in western cultures specifically in literature and movies views AI as something dangerous. Many stories evolve around how we lose control over artificial life (e.g Frankenstein's Monster, I Robot, 2001 A Space Odyssey) and how it consequently threatens their creators. Whereas in the Japanese culture artificial beings are often seen as positive forces that help humanity (e.g Astro Boy) and integrate well into their society.

Frederic Kaplan analyzed these differences and how they shaped the different perceptions of machines. His initial hypothesis can be seen in Table 1 from his paper [3]. (See copy of it below)

Table 1. Hypotheses about the differences in cultural acceptance of robots.

The West	Technology is central for defining what humans are	The possible convergence of humans and machines is a central topic, both fascinating and frightening	New robots can be upsetting
Japan	Technology has a more external role and can be part of an aesthetic quest	A distance is always maintained between the human body and technological prothesis	New robots rarely raise difficult issues

[1] A fact that has been acknowledged by researches and which they try to address by creating humanoid robots. [5]
[2] In an interview with PBS News Hour
[3] http://www-formal.stanford.edu/jmc/reviews/weizenbaum/node6.html#SECTION00060000000000000000
[4] http://profiles.nlm.nih.gov/ps/access/BBBBLN.pdf

In addition to the influence through literature and arts there is also a religious component that governs our view about artificial life. Even so religions do not play a direct role in the development of AI's they nevertheless shaped and still shape our views about life which indirectly influences our view of artificial life. Bartneck et al discuss this in [4] and one of the possible reason they mention is the different perception of the soul. Christianity sees the soul as something exclusive to humans whereas Confucianism and Buddhism also have the notion of souls for inanimate objects (rocks, waterfalls). They therefore believe that this could have influence the perception of artificial beings as something positive as well.

Machine Ethics and Roboethics

Even so there is a lot of discussion about ethics and AI the field of Machine Ethics and Roboethics are still very young. They try to scientifically address the ethical problems of artificial beings. Machine ethics is the idea[7] of explicitly adding an ethical dimension into machines. The field is concerned with the problem of creating machines that by themselves behave with certain ethical guidelines. Anderson and Anderson distinguish this as a separate problem from normal decision making because they argue that *"having all the information and facility in the world won't, by itself, generate ethical behavior in a machine"*[7].

Roboethics on the other hand is concerned with the problem of how humans should treat intelligent beings. Gianmarco Veruggio gives an overview [8] about the questions that drive Roboethics. (E.g. How far can we go in embodying ethics in a robot? Which kind of ethics is a robotic one?) The two fields try to address problems and questions that can at points interfere with each other, for example how the embedding of ethics into robots interferes with their autonomy.

References

[1] J. Weizenbaum, *Eliza : a computer program for the study of natural language communication between man and machine*. [Cambridge Mass.]: [MIT], 1965.

[2] J. Weizenbaum, *Computer power and human reason: From judgment to calculation.* .

[3] F. KAPLAN, "WHO IS AFRAID OF THE HUMANOID? INVESTIGATING CULTURAL DIFFERENCES IN THE ACCEPTANCE OF ROBOTS," *International Journal of Humanoid Robotics*, vol. 01, no. 03, pp. 465–480, Sep. 2004.

[4] C. Bartneck, T. Suzuki, T. Kanda, and T. Nomura, "The influence of people's culture and prior experiences with Aibo on their attitude towards robots," *AI & SOCIETY*, vol. 21, no. 1–2, pp. 217–230, May 2006.

[5] J. Schanze, *Plug and Pray*. 2010.

[6] B. Shneiderman, "Human Responsibility for Autonomous Agents," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 60–61, Mar. 2007.

[7] S. L. A. Michael Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine*, vol. 28, no. 4. p. 15, 15-Dec-2007.

[8] G. Veruggio, "The EURON Roboethics Roadmap," in *2006 6th IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 612–617.