Essentials of Robotics and Robot Laws

James H. Moor, one of the pioneering theoreticians in the field of computer ethics, defines four kinds of ethical robots.

- **Ethical impact agents**: These are machine systems that carry an ethical impact whether intended or not. At the same time, these agents have the potential to act unethical. Moor gives a hypothetical example called the 'Goodman agent', named after philosopher Nelson Goodman. The Goodman agent compares dates but has the millennium bug. This bug resulted from programmers who represented dates with only the last two digits of the year. So any dates beyond 2000 would be misleadingly treated as earlier than those in the late twentieth century. Thus the Goodman agent was an ethical impact agent before 2000, and an unethical impact agent thereafter.
- **Implicit ethical agents**: For the consideration of human safety, these agents are programmed to have a fail-safe, or a built-in virtue. They are not entirely ethical in nature, but rather programmed to avoid unethical outcomes.
- **Explicit ethical agents**: These are machines that are capable of processing scenarios and acting on ethical decisions. Machines which have algorithms to act ethically.
- **Full ethical agents**: These machines are similar to explicit ethical agents in being able to make ethical decisions. However, they also contain human metaphysical features. (i.e. have free will, consciousness, and intentionality)

Key questions:

➩ **The AI**-*control problem, The Superintelligence-Problem*
➩ Can we design genuine Artificial Moral Agents (AMAs) with responsibility, moral decision-making, autonomous moral evaluation?

International Symposium on Roboethics in 2004 by the collaborative effort of Scuola di Robotica, the Arts Lab of Scuola Superiore Sant'Anna, Pisa, and the Theological Institute of Pontificia Accademia della Santa Croce, Rome.

1. *Those who are not interested in ethics. They consider that their actions are* strictly technical, *and do not think they have a social or a moral responsibility in their work.*
2. *Those who are interested in* short-term ethical questions. *According to this profile, questions are expressed in terms of "good" or "bad," and refer to some cultural values. For instance, they feel that robots have to adhere to social conventions. This will include "respecting" and helping humans in diverse areas such as implementing laws or in helping elderly people. (Such considerations are important, but we have to remember that the values used to define the "bad" and the "good" are relative. They are the contemporary values of the industrialized countries).*
3. *Those who think in terms of* long-term ethical questions*, about, for example, the "Digital divide" between South and North, or young and elderly. They are aware of the gap between industrialized and poor countries, and wonder whether the former should not change their way of developing robotics to be more useful to the South. They do not formulate explicitly the question what for, but we can consider that it is implicit".*

(Gianmarco Verruggio, The Birth of Roboethics, 2005).

One of the pivotal questions in the field of *Applied Ethics* and *Digital Ethics* emerges:

⇨ Is something ethically desirable, only because it is technically possible?

Laws of Robotics

The "*Laws of Robotics"* are a set of fictional ethical guidelines created by science fiction writer Isaac Asimov. Asimov introduced these laws in his stories and novels to explore the relationship between humans and robots. The laws are as follows:

First Law

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

Second Law

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

Third Law

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

These laws were first explicitly stated in Asimov's 1942 short story "*Runaround*" and have since become a fundamental framework in science fiction literature and discussions about artificial intelligence and robot-ethics. Asimov later added a "Zeroth Law" in some of his later robot-related science fiction works.

Zeroth Law

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

It's important to note that these laws are fictional and do not exist in reality. In the real world, the field of robotics and artificial intelligence ethics is complex, and researchers and ethicists are working to establish ethical guidelines for the development and use of AI and robots.

MIRI (Machine Intelligence Research Intstitue), Berkeley, California

⇨ Friendly AI in the realm of *effective altruism*
⇨ positive for humanity and aligning with its values

Study questions

1) Why is it important to elaborate the implications of RE, ME, AIE in today's society?
2) Are there any significant loopholes or uncertainties in robot laws?
3) Is value alignment a too idealistic aim for the branch of Robotics?

Berthold