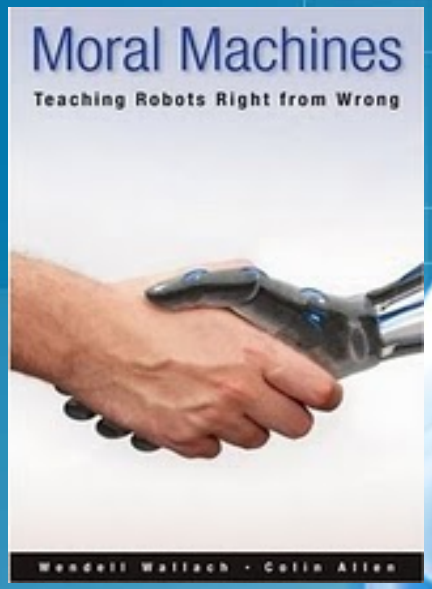


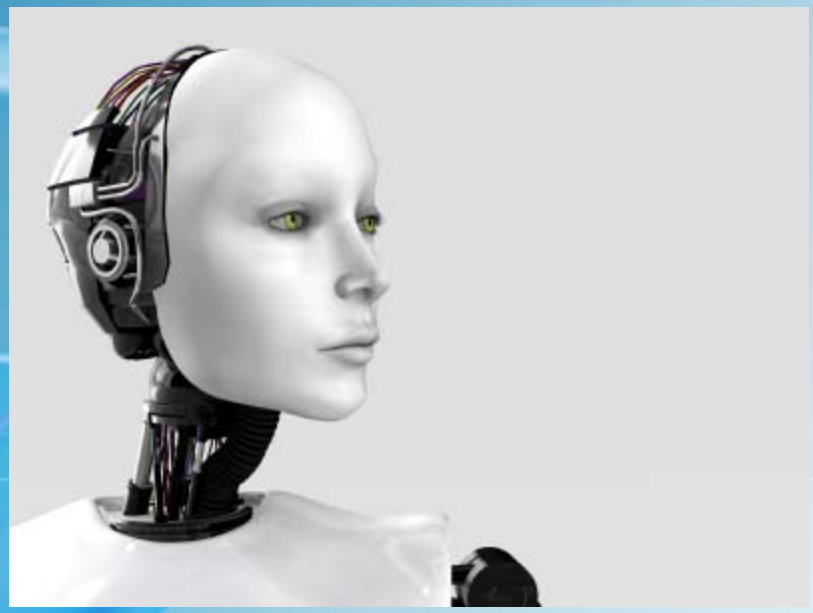
Moral Machines: Teaching Robots Right from Wrong

by

W. Wallach and C. Allen

Overview

- Introduction
- Why moral machines?
- Can robots be moral?
- Philosophers and Engineers about the design of AMAs
- Conclusions



Introduction

- (Ro)bot (physical robots and software agents): reality.
  - Driverless trains
  - Financial networks
  - Lethal weapons systems
  - Pets
  - Home appliances: Roomba!

- AMA (Artificial Moral Agent): dream?

No, say W. Wallach and C. Allen in their book

Why machine morality?

- Driverless trains: what should a good robot do?

Present-day examples of possible harm:

- System that control power grids
- Financial networks
- Lethal weapon systems
- Medical applications
- Reasons: faulty components, insufficient design

"The greater the freedom of a machine, the more it will need moral standards."

Can robots be moral?

The answer on this question is a list of other questions:

- What are the practical procedures for evaluating whether an action is right?
- How can a moral agent develop moral character?
- Will robots need emotions to function as adequate moral agents? When? How?
- The role of consciousness.

Philosophers and Engineers about the design of AMAs

Philosophers approach to decision making:

- Top-down. Use of rules, standards or theories to guide the design of a system's control architecture.

Engineers approach:

- Bottom-up. Rules are not explicitly defined, but the system learns about them through experience.

Conclusions

- Whether we want them or not they will appear, probably sooner than later
- Why? It's all about power.
- Humans are moral agents, but are they always in reality?
- So who or what to fear more: humans or machines?