Autonomous Machines Are Ethical

John Hooker

Carnegie Mellon University

INFORMS 2017

Thesis

- Concepts of deontological ethics are **ready-made** for the age of AI.
  - Philosophical concept of **autonomy** applies immediately to robot ethics.

Thesis

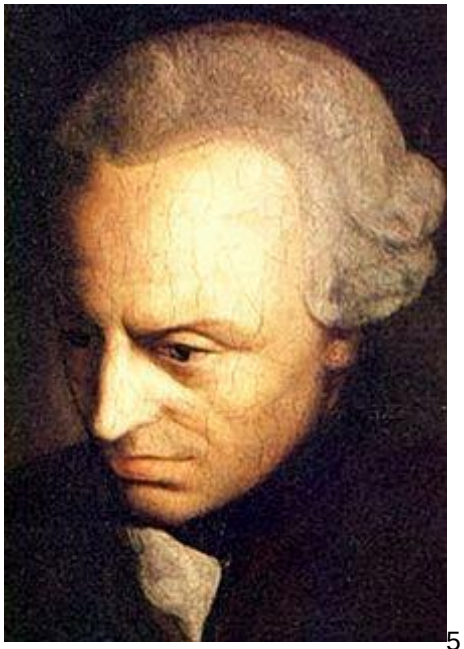
- Concepts of deontological ethics are **ready-made** for the age of AI.
  - Philosophical concept of **autonomy** applies immediately to robot ethics.
  - One conclusion: **autonomous** machines are **ethical**.
  - Other basic issues resolved.

Autonomy

- Popular sense:
  - Autonomous = **Self-controlling**; not directly controlled by another agent.

Autonomy

- A deeper philosophical sense:
  - Autonomous = Can be explained by **reasons** adduced by the agent.
  - Even while **also** explicable as the result of physical and biological causes.
  - "**Dual standpoint**" theory.



Immanuel Kant

Deontological Ethics

- Unethical = **no coherent rationale**.
  - Unethical behavior is not really **action**.
  - Ethics = an imperative to exercise **agency**.
  - Underlying premise: **universality of reason**.
    - Reasons that justify an action for **one agent** justify the action for **any agent** to whom the reasons apply.

Generalization Principle

- An action and its rationale should be **generalizable**.
  - It must be rational to believe that the **reasons** for an action are consistent with the assumption that **all agents** who have the same reasons act the same way
    - …where the reasons have maximal **scope.**
    - Otherwise, the agent sees the reasons as justifying the act and **not** justifying the act.

Generalization Principle

- Example: **lying** merely because it is **convenient** for others to **believe** the lie.
  - No one would **believe** lies if **everyone** who found it convenient to lie did so.
  - So lying merely for convenience is **not generalizable**.

Generalization Principle

- Other examples of **ungeneralizable** behavior.
  - **Breaking a contract** to save money.
  - **Cheating** on an exam to get a better job.
  - **Breaking a promise** merely because one doesn't want to keep it.

Respect for Autonomy

- An action can be regarded as a conditional **action plan**.
  - "If reasons A, B and C apply, then do X."
  - Example "If I want to catch a bus, and there is a bus stop across the street, and no cars are coming, then I will cross the street."

Respect for Autonomy

- "If I want to catch a bus, and there is a bus stop across the street, and no cars are coming, then I will cross the street."
  - **Violation** of my autonomy: you pull me out of the street as I cross.
  - **Not** a violation of autonomy: you pull me out of the path of an oncoming car.

Respect for Autonomy

- **Joint autonomy principle**
  - My action plan must not **interfere** with the joint execution of the (ethical) action plans of other agents.
    - **…unless** there is informed or implied consent.
  - **Why?** Universality of reason.
    - I could be one of the other agents.

Utilitarian Principle

- Can be viewed as a **deontological** principle.
  - Utility = what I regard as **intrinsically valuable** (e.g., happiness)

Utilitarian Principle

- Can be viewed as a **deontological** principle.
  - Utility = what I regard as **intrinsically valuable** (e.g., happiness)
  - **Principle**: I should choose an act that that I can rationally regard as **maximizing** the net expected utility of all agents affected.
    - …where **only acts** that satisfy the generalization and autonomy principles are considered.

Machines as Agents

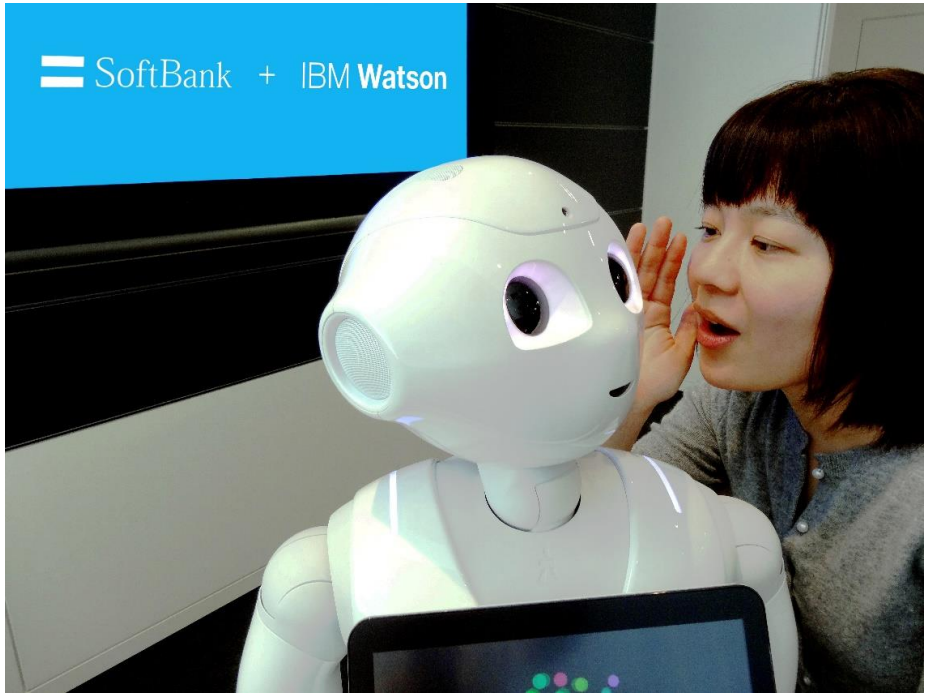
- A **machine** is an **agent** if it is capable of adducing reasons for its actions.
  - For example, household robot.

Machines as Agents

- A **machine** is an **agent** if it is capable of adducing reasons for its actions.
  - For example, household robot.
  - This does **not** anthropomorphize machines.
    - An agent need not be a human agent.

Duties to Machines

- Actions toward machines must be **generalizable**.
  - Should not lie to your robot.

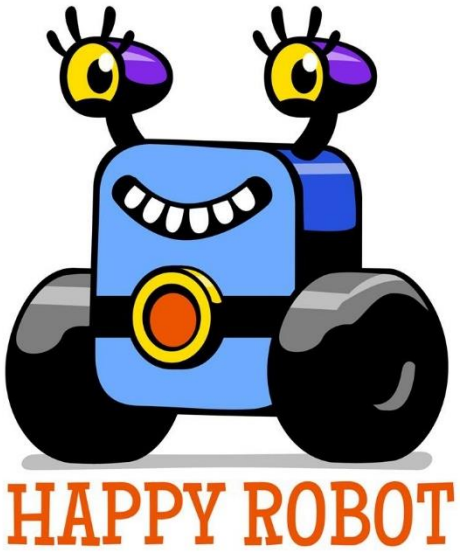


Duties to Machines

- Respect machine **autonomy**.
  - Should not throw obsolete machine in the trash.
  - What if machines are immortal due to replacement parts? Overpopulation problem?

Duties to Machines

- Not clear that we have **utilitarian** obligations to machines.
  - Human-oriented utility (e.g. happiness) may not apply to non-sentient machines.



Duties of Machines

- Machine's actions should be **generalizable**.
  - Argument for the generalization principle presupposes only formal properties of agency, not humanity.

Duties of Machines

- Machine's actions should be **generalizable**.
  - Argument for the generalization principle presupposes only formal properties of agency, not humanity.
- Machines should respect **autonomy**.
  - Ditto.

Duties of Machines

- Machine's actions should be **generalizable**.
  - Argument for the generalization principle presupposes only formal properties of agency, not humanity.
- Machines should respect **autonomy**.
  - Ditto.
- **Utilitarian** obligations?
  - Perhaps not.

Duties of Machines

- So autonomous machines are **ethical**.
  - At least with respect to generalization and autonomy principles.

Robot Masters?

- Will superintelligent, autonomous machines **take over**?

Robot Masters?

- Will superintelligent, autonomous machines **take over**?
- **No!** This violates human autonomy

Robot Masters?

- Will superintelligent, autonomous machines **take over**?
- **No!** This violates human autonomy.
  - Autonomous machines will not **reprogram** themselves to be unethical.
    - This is unethical!

Responsibility

- Should **machines** be held **responsible** for their actions?
  - Or their **human** designers?

Responsibility

- Should **machines** be held **responsible** for their actions?
  - Or their **human** designers?
- **Neither.**
  - Unethical behavior is **never freely chosen**, because it is not action.
  - So agents are never “responsible” for their unethical behavior in the ordinary sense.

Responsibility

- However, we can encourage acts for which agents can give coherent reasons.
  - This is consistent with physical determinism, and in fact **requires** it.

Responsibility

- However, we can encourage acts for which agents can give coherent reasons.
  - This is consistent with physical determinism, and in fact **requires** it.
  - How to do this?
    - Training.
    - Punishment and reward.
    - Ethics instruction.
    - None of this presupposes that agents are “responsible” for their actions.

Living with Machines

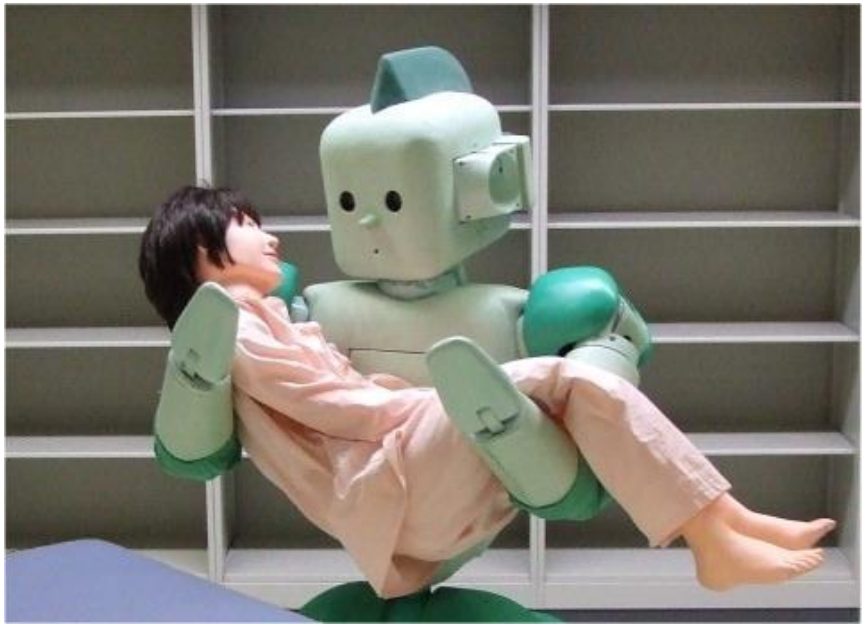
- It may be easier to teach ethics to machines than to people.
  - Maybe it's not so bad to have a **fully ethical** segment of the population.

Living with Machines

- What if machines have no **utilitarian** obligations to us?
  - They don’t care about our happiness, etc.

Living with Machines

- We can build machines that **prefer** human happiness.



Living with Machines

- We can build machines that **prefer** human happiness.
  - Determining preferences is **consistent** with agency.
    - After all, **human** preferences/culture are largely determined by external factors.
    - But we must make sure machines don't **reprogram** their preferences.

Discussion?