Robots as Agents

Module 12 of a course on *Ethical Issues in AI*

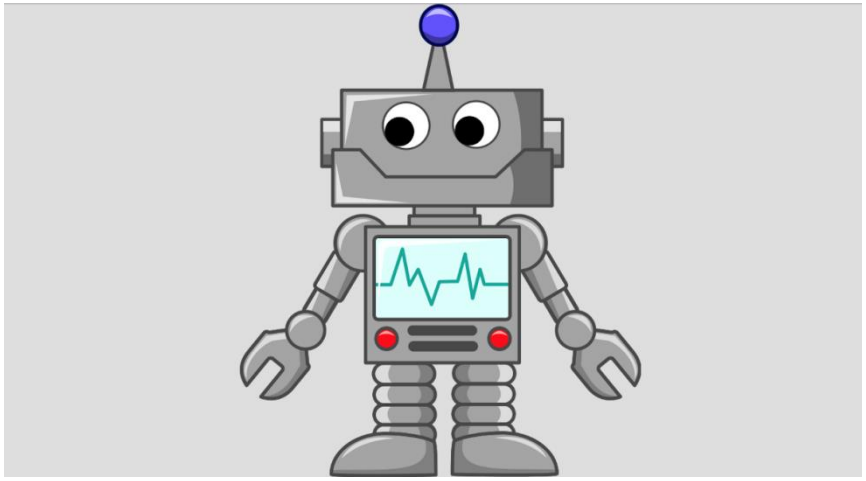
Prepared by

John Hooker
Emeritus Professor, Carnegie Mellon University

CMU Osher, February 2025

Autonomous robots

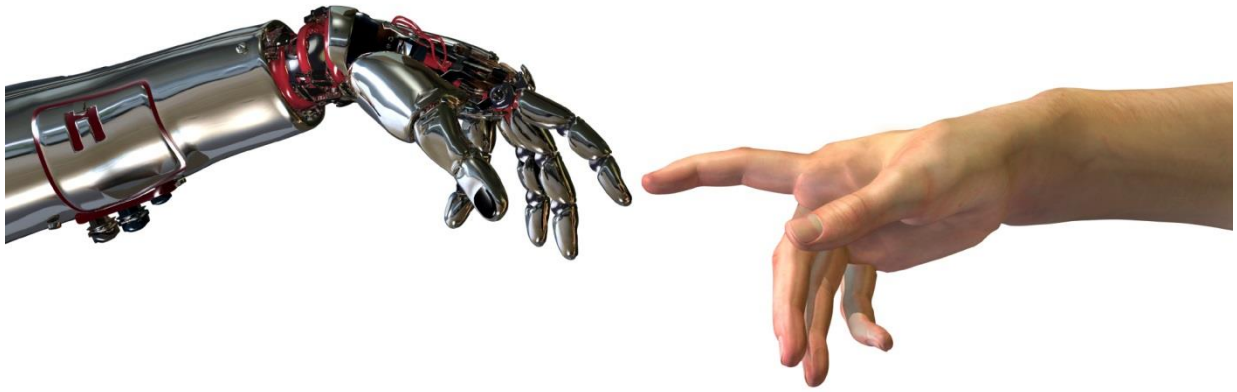
- Are autonomous robots **responsible** for their actions?
  - *Do they have **obligations**?*
  - *Do **we** have obligations to machines?*



Autonomous robots

- What about **superintelligent** machines?
  - *…after a technological "singularity"?*

    Vernor Vinge, *The Coming Technological Singularity,* 1993.
    - Machines will reprogram themselves.
    - Will they take over?



Autonomous robots

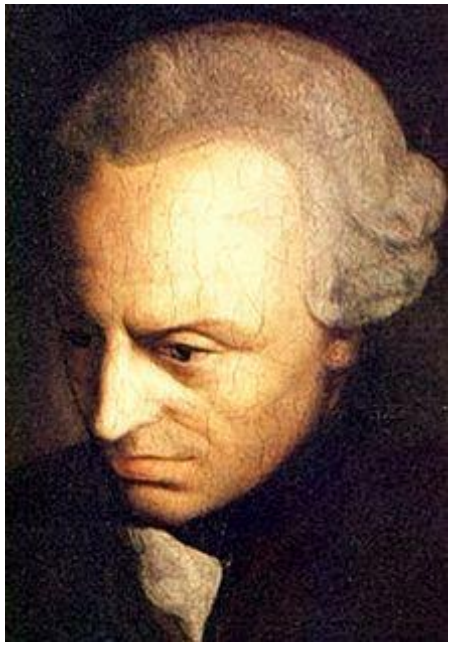
- Concepts of deontological ethics are **ready-made** for the age of AI.
  - *Concept of **autonomy** applies immediately to robot ethics.*
  - *One conclusion: **truly autonomous** machines are **ethical**.*

Autonomy

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

Autonomy

- The deeper philosophical sense we use:
  - *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

Autonomy

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

Autonomy

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

Duties TO machines

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

Duties TO machines

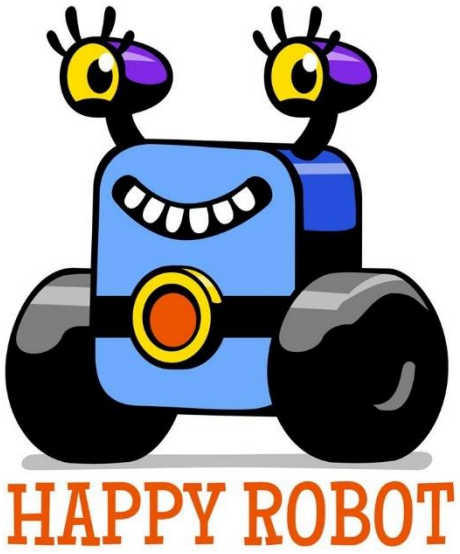
- Respect machine **autonomy**.
  - *Should not throw obsolete machines in the trash?*
    - What if machines are **immortal** due to replacement parts? Overpopulation problem?

Duties TO machines

- Respect machine **autonomy**.
  - *Should not throw obsolete machines in the trash?*
    - What if machines are **immortal** due to replacement parts? Overpopulation problem?
    - Solution: Build robots that **want to die**…
    - …much as nature builds humans who **want to live**.

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 the world**?

Robot masters?

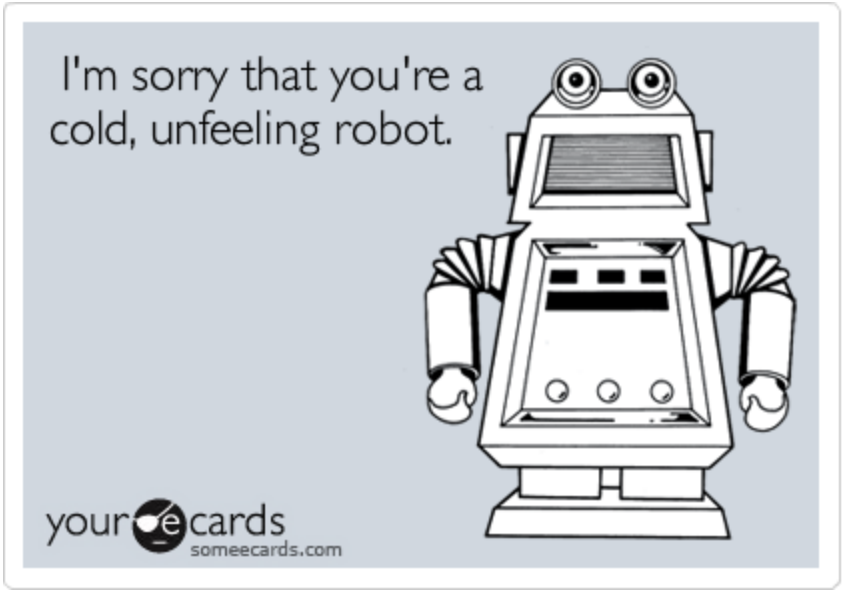
- Will superintelligent, autonomous machines **take over the world**?
- **No!** This violates human autonomy.

Robot masters?

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

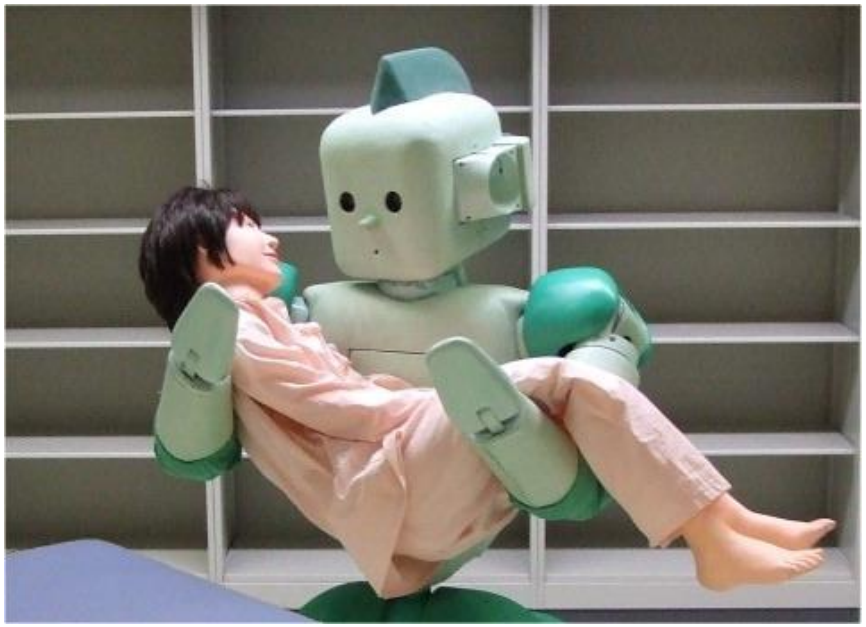
Living with machines

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



Living with machines

- We can build machines that are hardwired to **prefer human happiness**.



Living with machines

- Building autonomous machines may be a **bad idea**.
  - We may **fail** to make them autonomous!
- But… it may be **easier** to teach ethics to machines than people.