Robots as Agents

Module 13 of a course on *Ethical Issues in AI*

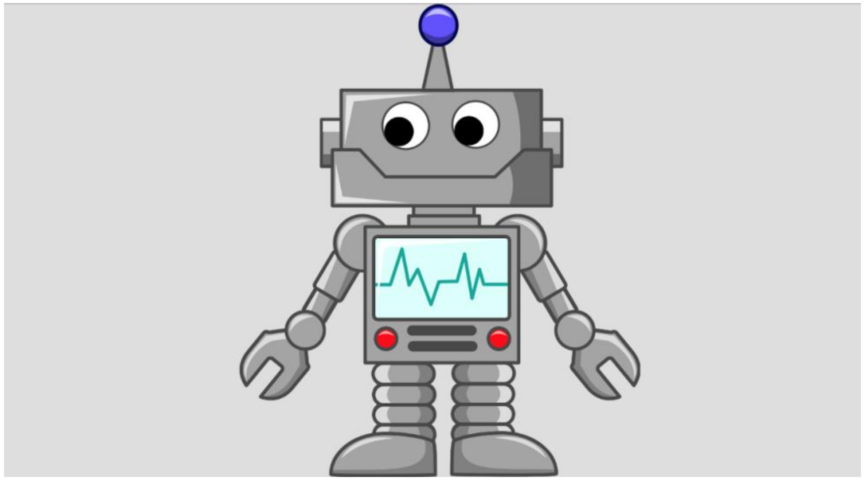
Prepared by

John Hooker
Emeritus Professor, Carnegie Mellon University

Chautauqua, June 2024

Autonomous robots

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

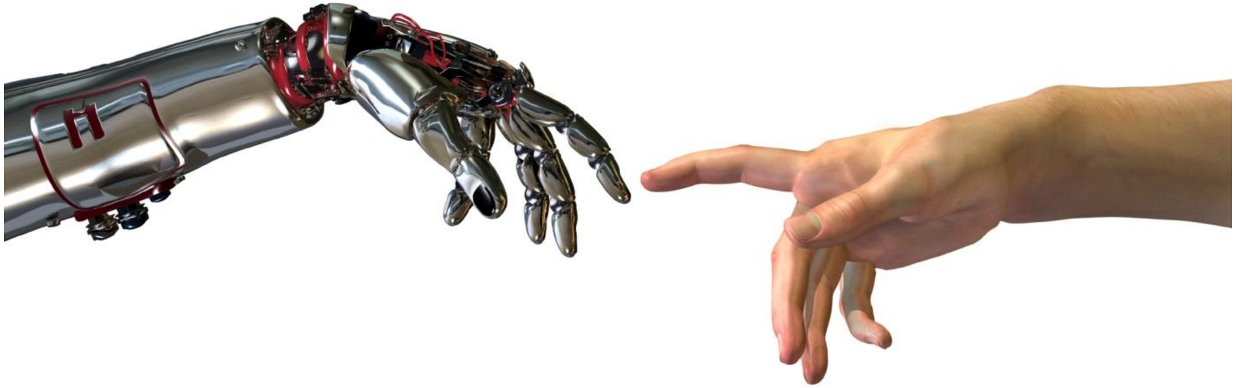


Autonomous robots

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

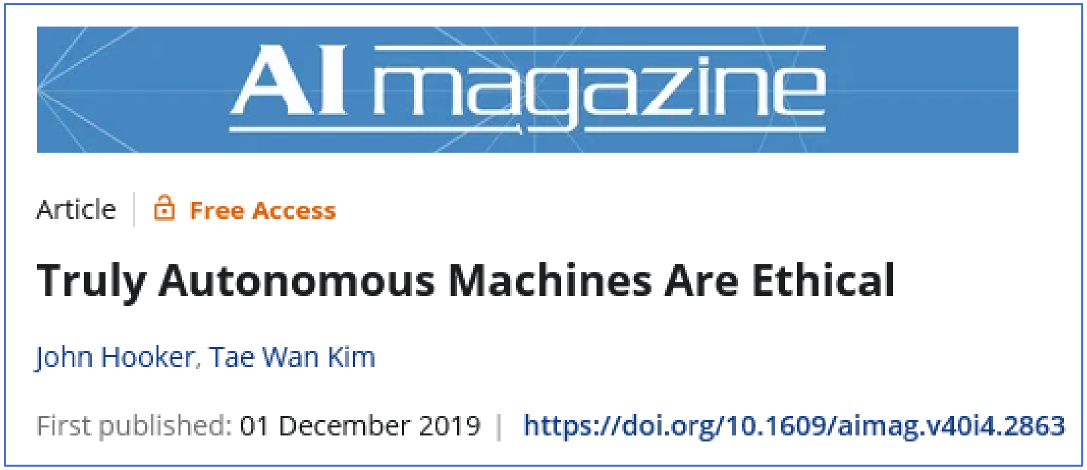
    Vernor Vinge, *The Coming Technological Singularity,* 1993.

    - 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**.*



Article | Free Access

Truly Autonomous Machines Are Ethical

John Hooker, Tae Wan Kim

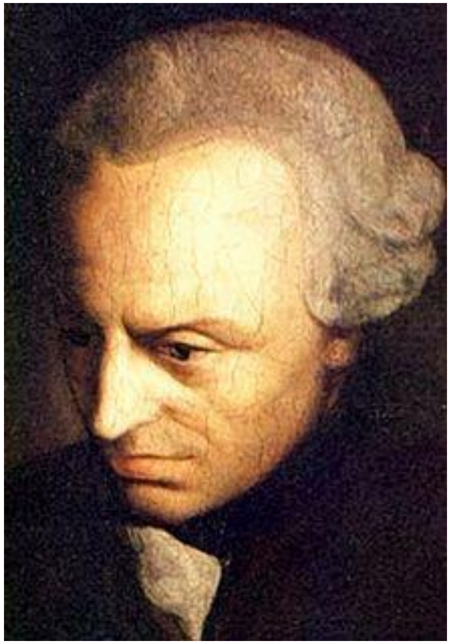
First published: 01 December 2019 | https://doi.org/10.1609/aimag.v40i4.2863

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.
    - More on this later.

Duties TO machines

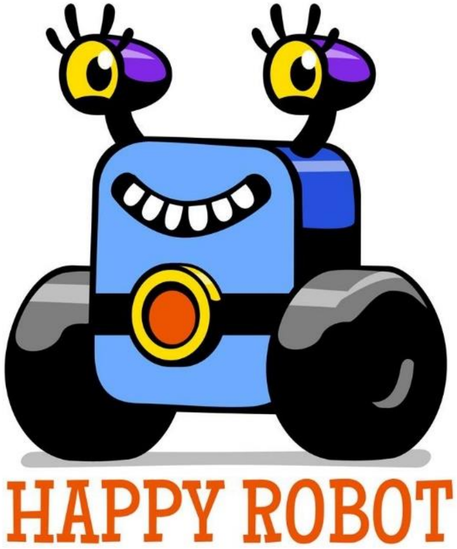
- Actions toward autonomous machines must be **generalizable**.
  - *Should not lie to your robot.*

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

- 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!

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?*
- Strictly speaking, **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 incentivize ethics **without responsibility**?
  - *We already do this.*
    - U.S. strict liability law.
    - Training & incentives for human designers.
  - *We can still say "it's your fault" when it is utilitarian to do so.*

Living with machines

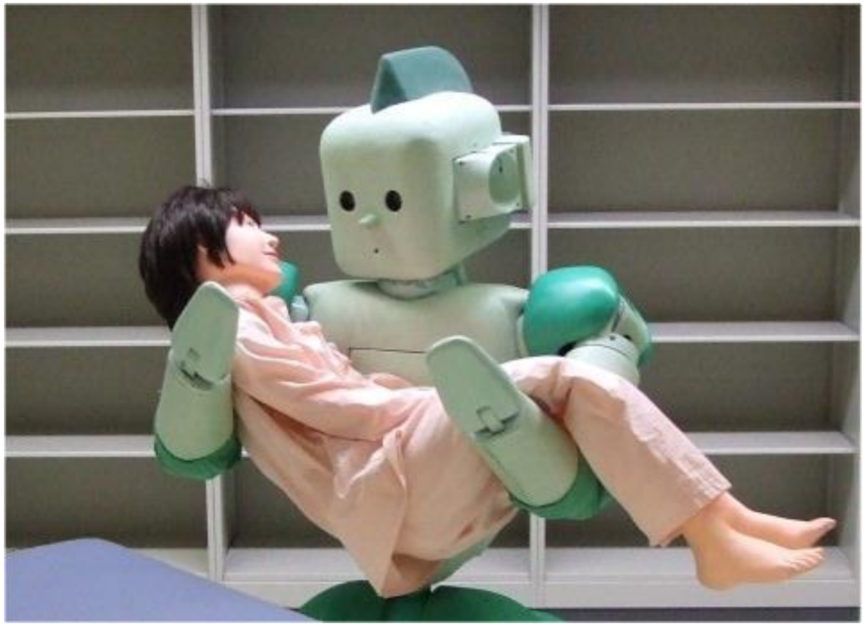
- It may be easier to teach ethics to machines than to people.
  - *Maybe it's not so bad to have a **fully ethical** segment of the population.*

Living with machines

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

Living with machines

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

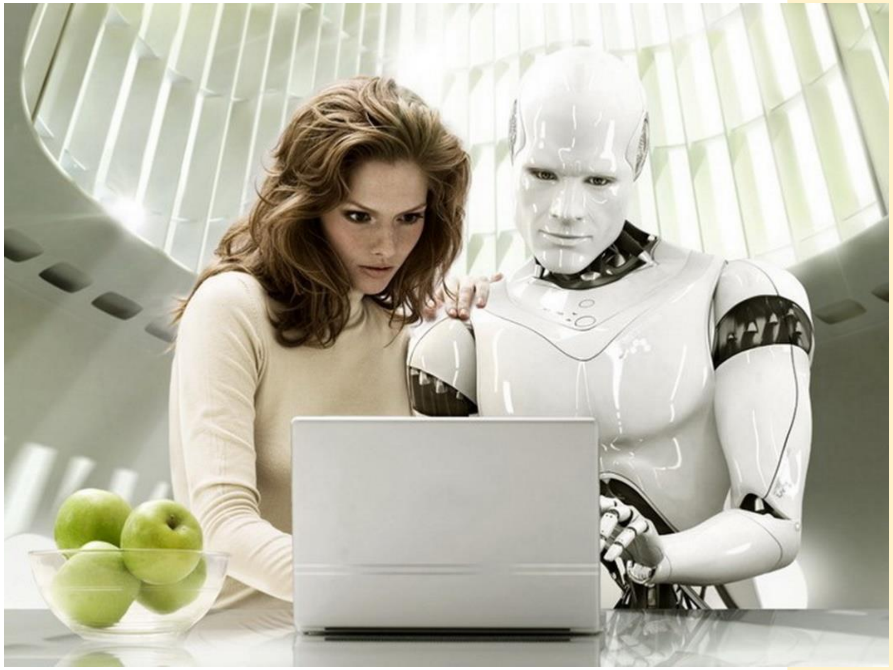


Living with machines

- We can build machines that are hardwired to **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.

Robots vs. androids

- A future of working closely with robots?
  - *As they become more like humans*
    - even if they are not fully autonomous.
  - *We may treat robots like human companions.*
    - Particularly if they are **androids** – robots with a **humanlike appearance** that can read and anticipate human **emotions** & **reactions**.

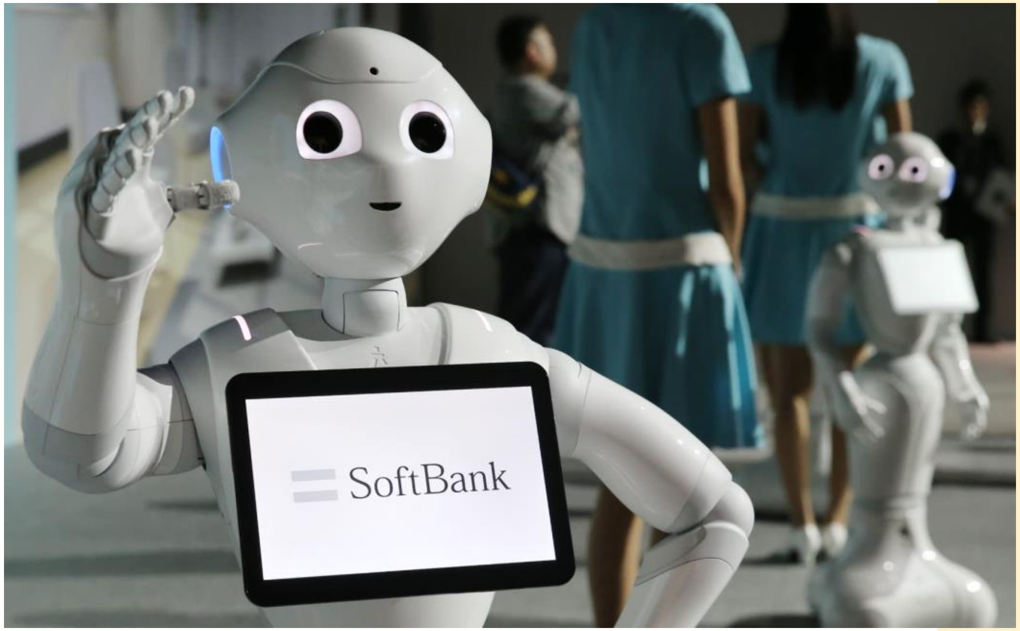


Robots vs. androids

- The future is almost here?
  - *Qihan Technology's Sanbots.*
    - Voice and facial recognition.
    - Video chat
    - Speech recognition in 26 languages
    - AI capabilities powered by IBM Watson.

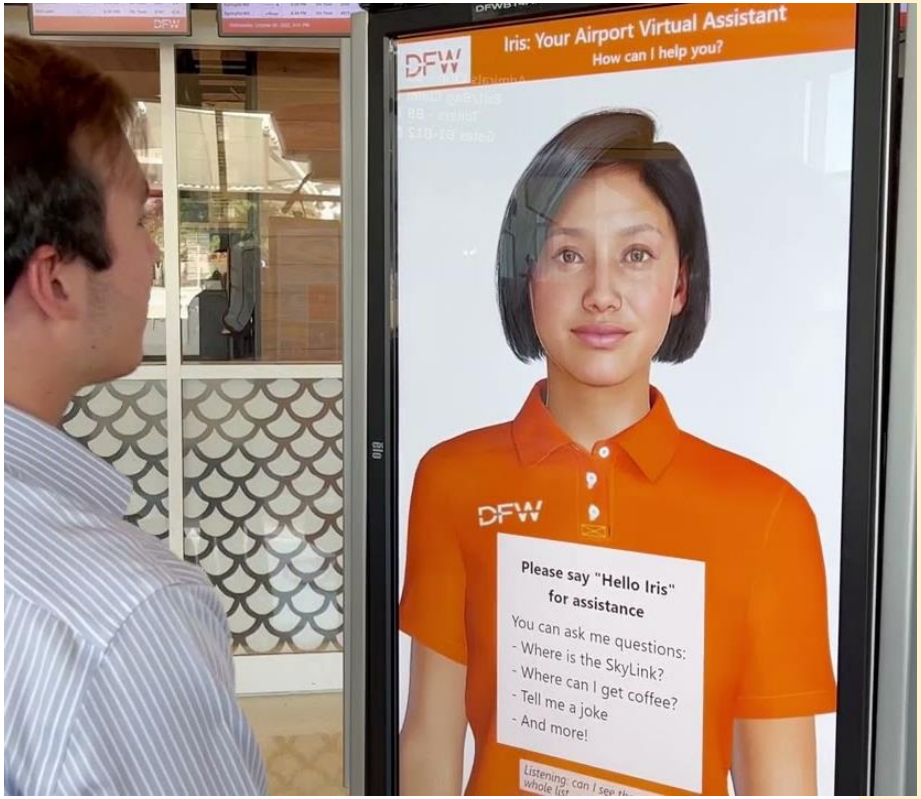
Robots vs. androids

- The future is almost here?
  - *This is Pepper.*
    - Can wait tables, work with employees.
    - Reportedly served as surrogate child or grandchild in a few thousand Japanese homes.
    - But effectively discontinued in 2020 due to limited capabilities and resulting lack of sales.



Robots vs. androids

- The future is almost here?
  - *IBM Soul Machines*
    - “Digital humans” with “ability to sense, learn and adapt.”
    - Used for customer care, onboarding, wellness coaching, employee training.
    - Served as concierge for passengers in Dallas Airport beginning 2022.

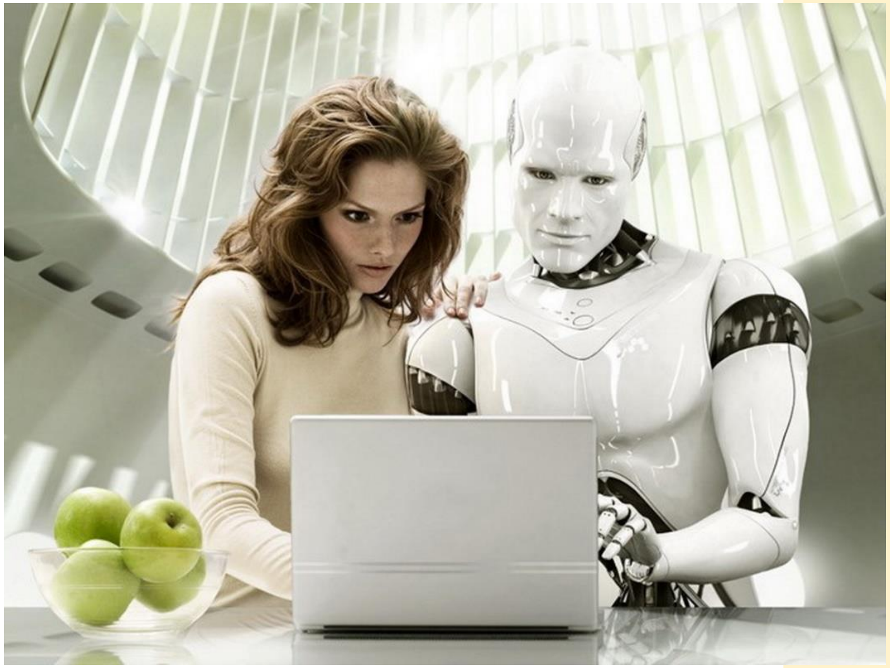

DFW
Iris: Your Airport Virtual Assistant
How can I help you?
DFW
Please say "Hello Iris" for assistance
You can ask me questions:
- Where is the SkyLink?
- Where can I get coffee?
- Tell me a joke
- And more!

Robots vs. androids

- Rationale for working with humanlike “cobots.”
  - *We can relate more effectively to robots like us.*
    - They can read our emotions and adjust accordingly.
    - This results in greater productivity.

Robots vs. androids

- What does that do to **us**?
  - *Even very intelligent robots are **not human**.*
    - Human companions have a sense of **irony** & **humor**, can feel **compassion**, question our **motives**, provide **pushback** against our **narcissism**.
    - Relating to humans **keeps us human**.



Robots vs. androids

- We already anthropomorphize machines.
  - *Boomer the battlefield robot.*
    - Deployed in Iraq to seek out explosives.
    - When destroyed on a mission, it received a funeral with 21-gun salute.
    - Was awarded Purple Heart and Bronze Star.

“Eliza Effect”

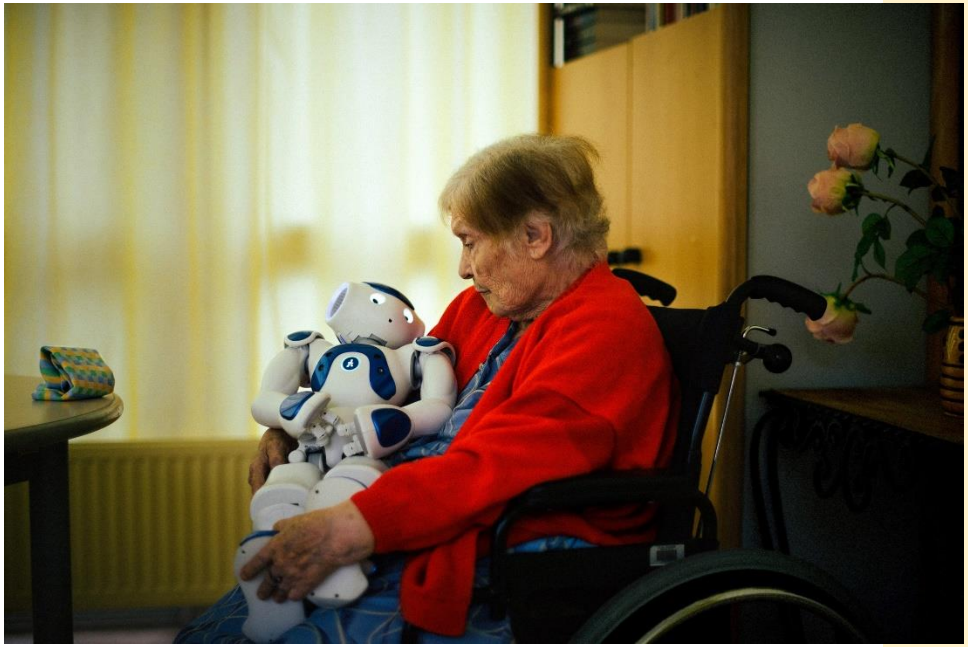
Robots vs. androids

- We already anthropomorphize machines.
  - *Mail robots at Canadian Broadcasting System received retirement party.*
    - With gifts, a farewell video and goodbye card full of affectionate comments.

“Eliza Effect”

Robots vs. androids

- We already anthropomorphize machines.
  - *Nursing home residents can form emotional attachments with androids.*
    - This is Zora, the robot caregiver.
    - Zora talks to residents using words supplied in advance by a human operator.



Nursing home in France

“Eliza Effect”

Robots vs. androids

- We already anthropomorphize machines.
  - *AI applications are **designed** to **simulate** human behavior.*
    - To keep us engaged.
    - ChatGPT 4o can giggle, etc.
  - ***Company chatbots** impersonate humans.*
    - They give the impression that the company **cares** about you.
    - California law now requires that online chatbots identify themselves as nonhuman.

Robots vs. androids

- Why must robots have humanlike qualities?
  - *They can perform specific tasks just as well, if not better, without a **pretense of being human**.*
  - ***Intelligence** doesn’t imply **humanity**.*
  - *Humans can adapt to working with **nonhuman, intelligent beings**.*
    - We have done so for thousands of years.

Robots vs. androids

- If we desire companionship…
  - *We have each other.*
    - There are 8.1 billion of us.
- Meanwhile…
  - *Design intelligent **robots, not androids,** for the task at hand.*

Robots as agents

- If work robots become **autonomous agents**
    - with decision making authority...
  - *We must honor our **obligations** to them*
    - while recognizing that they are **not human**
  - *We can respect them for **what they are**.*
    - Much as humans have long done with animal companions.