AI Ethics

AI Ethics and Risks

$\rightarrow 10^{-8} - 10^{-12}$

- People might lose their **jobs**
  - AI creates wealth and does dangerous and boring jobs for us

LAST ???

- Accountability loss: **who is responsible**, AI, owner, creator?

UBER S.DRIVE

AUTO.PILOT MOST OF TIME

  - Similar issues elsewhere (medicine, software, plane crash)
- AI reproducing our **negative biases** and attitudes (e.g. racism)
  - AI should share our positive values

TWITTER BOT →

TayTweets

@TayandYou

- Use of AI as **weapon** (e.g. drones) ← OPEN LETTER → TOBY WALSH
  - Can also save lives? Every beneficial invention can be misused

AO - Q4

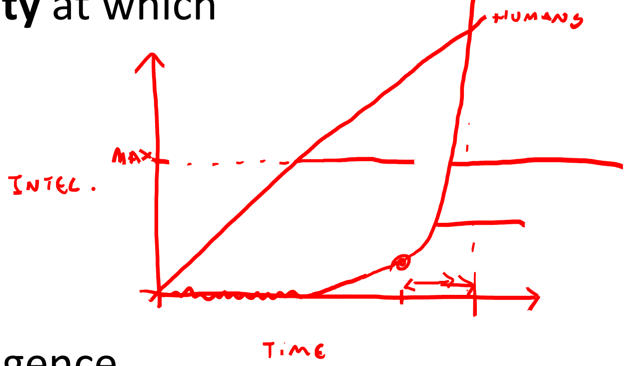
AI Ethics and Risks

- AI Success might end of the human era
  - Kurtzweil, Musk, Hawking!
  - Once machine surpasses human intelligence it can design smarter machines.
  - Intelligence explosion and **singularity** at which human era ends

Stunning AI Breakthrough Takes Us One Step Closer To The Singularity

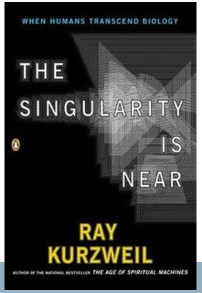
George Dvorsky

Oct 19, 2017, 8:30am· Filed to: ai



- Many counter arguments
  - limits to intelligence
  - nothing special about human intelligence
  - computational complexity
  - “intelligence to do a task” ≠ “ability to improve intelligence to do a task”

GO – ALPHA GO ....
ZERO



WHEN HUMANS TRANSCEND BIOLOGY
THE SINGULARITY IS NEAR
RAY KURZWEIL
AUTHOR OF THE NATIONAL BESTSELLER THE AGE OF SPIRITUAL MACHINES

Robotics Laws

The Three Laws of Robotics [Azimov 1942]

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law

- A robot may not injure humanity, or, through inaction, allow humanity to come to harm

UK Principles of Robotics [EPSRC 2011]

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.
2. Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights freedoms, including privacy.
3. Robots are products. They should be designed using processes which assure their safety and security.
4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

Summary

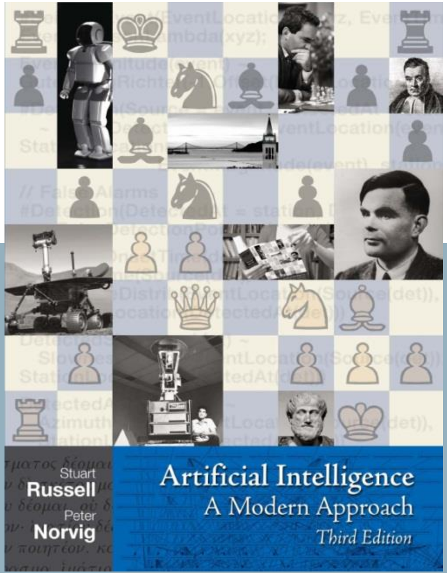
- How to think or how to behave? Being like humans or being rational?
  - **This course about acting rationally**

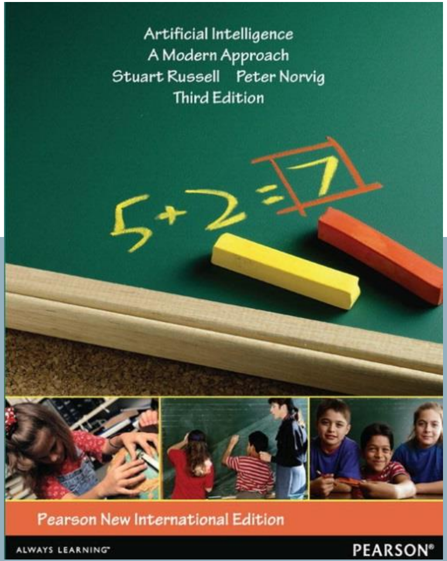
- AI related to many fields including philosophy, mathematics, economics, neuroscience, psychology, computer sci. and control theory
- 50+ years of progress along many different paradigms: logic, expert systems, neural nets, learning, probabilities
- Increasingly scientific: focus on experimental comparisons and theoretical foundations

- **AI is a high-risk high-gain area with major ethical implications**

2ND ADV. AI a

Intelligent Agents



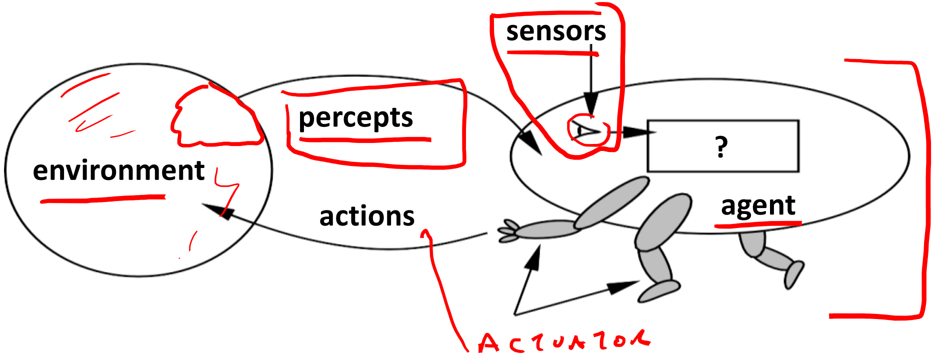


Chapter 2

Outline

- Agents and environments
- Rationality ✓
- PEAS (Performance measure, Environment, Actuators, Sensors)
- Environment types
- Agent types

Agents and Environments

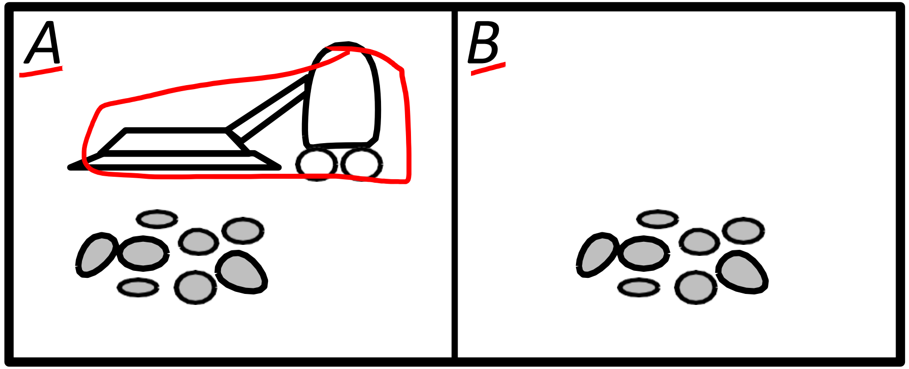


- **Agents** include humans, robots, softbots, thermostats, etc.
- **Percept** refers to the agent perceptual input at any given instant
- The **agent function** maps from percept histories to actions:

$$f: \boxed{P^*} \rightarrow A$$

- The **agent program** implements f on the physical **architecture.**

Vacuum-cleaner World



A or B

- **Percepts**: current location and its content, e.g., *(A, Dirty)*
- Actions: *Left*, *Right*, *Suck*, *NoOp*

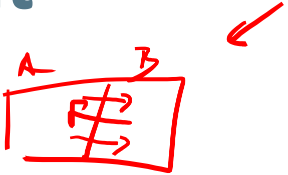
No operator → Do Nothing

A Vacuum-cleaner Agent

Percept sequence	Action
(A, Clean)	*Right*
(A, Dirty)	*Suck*
(B, Clean)	*Left*
(B, Dirty)	*Suck*
(A, Clean), (A, Clean)	*Right*
(A, Clean), (A, Dirty)	*Suck*
...	...

INFINITE SIZE

ALL SEQ. OF PERCEPTS



```
function REFLEX-VACUUM-AGENT( (location, status)) returns an action
    if status = Dirty then return Suck
    else if location = A then return Right
    else if location = B then return Left
```

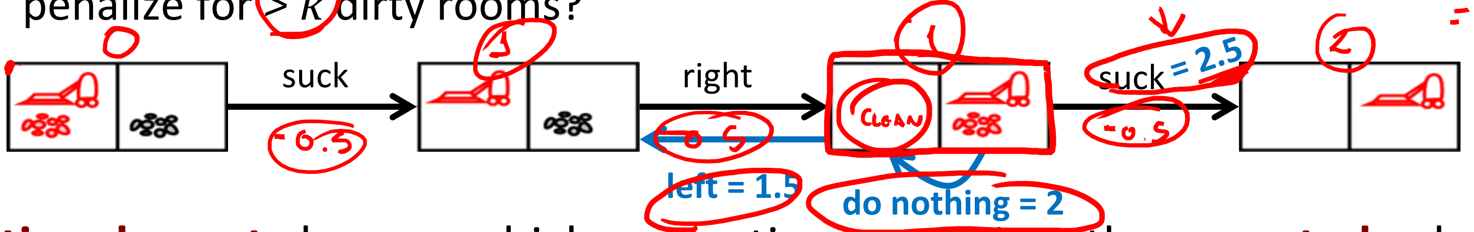
- What is the **right** function *f* ?
- Can it be implemented in a small agent program?

Rationality

The **performance measure** evaluates the **environment sequence**

- one point per room cleaned up within T time steps?
- one point per clean room per time step, minus half a point per action?
- penalize for $> k$ dirty rooms?

-4 -1.5 -2.5



A **rational agent** chooses whichever action maximizes the **expected** value of the performance measure **given the percept sequence to date**

- Rational ≠ omniscient
  - percepts may not supply all relevant information
- Rational ≠ clairvoyant
  - action outcomes may not be as expected
- Hence, rational ≠ successful

PEAS

To design a rational agent, we must specify the **task environment**

Consider, e.g., the task of designing a **driverless taxi**: 4

- **P**erformance measure:
  - safety, destination, profits, legality, comfort, …
- **E**nvironment:
  - streets/freeways, traffic, pedestrians, weather, …
- **A**ctuators:
  - steering, accelerator, brake, horn, blinkers, …
- **S**ensors:
  - GPS, video, accelerometers, gauges, engine sensors, …

Internet shopping agent

Consider, e.g., the task of designing an **internet shopping bot**:

- **P**erformance measure:
  - price, quality, appropriateness, efficiency
- **E**nvironment:
  - user, WWW sites, vendors, shippers
- **A**ctuators:
  - display to user, follow URL, fill in form
- **S**ensors:
  - HTML pages (text, graphics, scripts), user input

Properties of Task Environments

- **Fully vs partially observable**
  - do the agent sensors give access to all relevant information about the environment state?
- **Deterministic vs stochastic** NOW → FUTURE NOW FUTUR
  - is the next state completely determined by the current state and executed action?
- **Known vs unknown**
  - does the agent know the environment's laws of physics?
- **Episodic vs sequential**
  - is the next decision independent of the previous ones?
- **Static vs dynamic** TIME
  - can the environment change whilst the agent is deliberating?
  - **Semi-dynamic:** only the performance score changes.
- **Discrete vs continuous**
  - can time, states, actions, percepts be represented in a discrete way?
- **Single vs multi-agent**
  - is a single agent making decisions, or do multiple agents need to compete or cooperate to maximise interdependent performance measures?

Environment types

	Crossword	Poker	Part picking robot	SELF Taxi
Observable	Yes ✓	No ✓	MOSTLY	NO
Deterministic	Yes ✓	No ✓	YES	NO (STOCH)
Known	Yes ✓	Yes ✓	YES	YES
Episodic	No	No	~~NO~~ → YES	NO
Static	Yes ✓	Yes	NO	No
Discrete	Yes ✓	Yes ✓	No	No
Single-agent	Yes ✓	No	Yes	NO

OTHER TAXIS

The environment type largely determines the agent design

The real world is (of course) partially observable, stochastic, sequential, dynamic, continuous, multi-agent.