TDT4171 Artificial Intelligence Methods

Lecture 1 – Introduction and Foundation

Norwegian University of Science and Technology

Helge Langseth
Gamle Fysikk 255
helge.langseth@ntnu.no

NTNU

Outline

1. Practical information
   - About TDT4171
   - The other stuff
2. Philosophical Foundations – and movie bonanza
   - What is *Intelligence*?
   - The Touring Test
   - The Chinese Room
   - Ethics

TDT4171

Goals of the course:

"[. . .] The three main ways of reasoning (rule-based, model-based, and case-based), will be discussed, with most focus given to model-based reasoning. In particular, we work with reasoning with uncertain and/or partly missing information, as well as the basis for learning systems (machine learning)."

Expected background:

- TDT4109 Introto Information Technology: Python programming
- TMA4240 Statistics: Manipulation of probabilities
- TDT4172 Intro to ML: Basic ML understanding.
  No worries. Backup solution available.

TDT4171

Goals of the course:

"[. . .] The three main ways of reasoning (rule-based, model-based, and case-based), will be discussed, with most focus given to model-based reasoning. In particular, we work with reasoning with uncertain and/or partly missing information, as well as the basis for learning systems (machine learning)."

Syllabus (Tentative, as of Jan 17, 2025):

- The text-book *"Artificial Intelligence – A Modern Approach"* by S. Russell and P. Norvig, **4th ed.**, 2020. Approx. 300p.
- A. Aamodt and E. Plaza (1994): *"Case-based reasoning; Foundational issues, methodological variations, and system approaches"*. 20p.
- D. Ganesan and S. Chakraborti (2018): *"An Empirical Study of Knowledge Tradeoffs in Case-Based Reasoning*. 7p.
- Also: *"Deep Learning"* by I. Goodfellow, Y. Bengio and A. Courville, 2016. Selected parts as supporting materials.

The grand vision

An autonomous self-moving machine that **acts**, **reasons**, and **learns** intelligently.

The grand vision

An autonomous self-moving machine that **acts**, **reasons**, and **learns** intelligently.

The AI hype

Marvin Minsky on the future of AI — in 1970!

. . . in three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, and have a fight. At that point, the machine will start to educate itself with fantastic speed. In a few months it will be at genius level and a few months after that its powers will be incalculable.

The AI hype

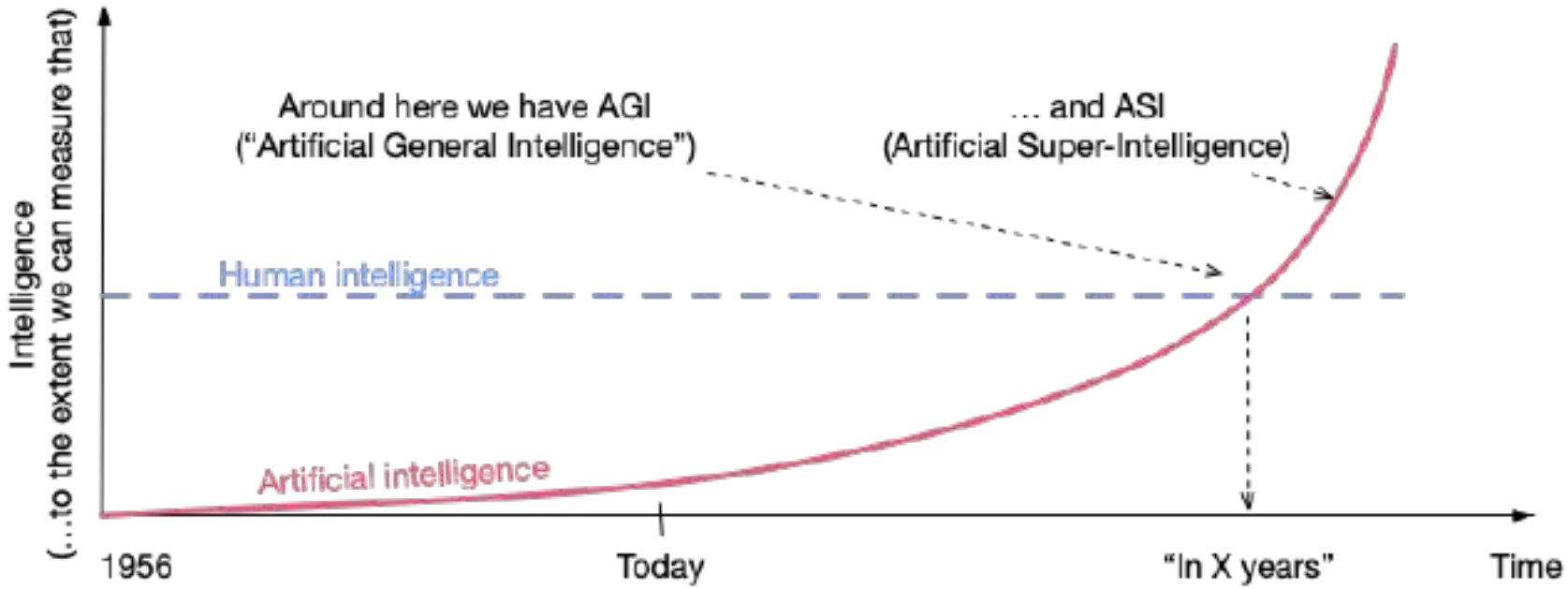
Marvin Minsky on the future of AI — in 1970!

. . . in three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, and have a fight. At that point, the machine will start to educate itself with fantastic speed. In a few months it will be at genius level and a few months after that its powers will be incalculable.

More realistic version — Minsky in 2013

We have machines, which can solve problems and/or interact with users operating in a clearly defined environment. There is no reading of Shakespeare. . . yet.

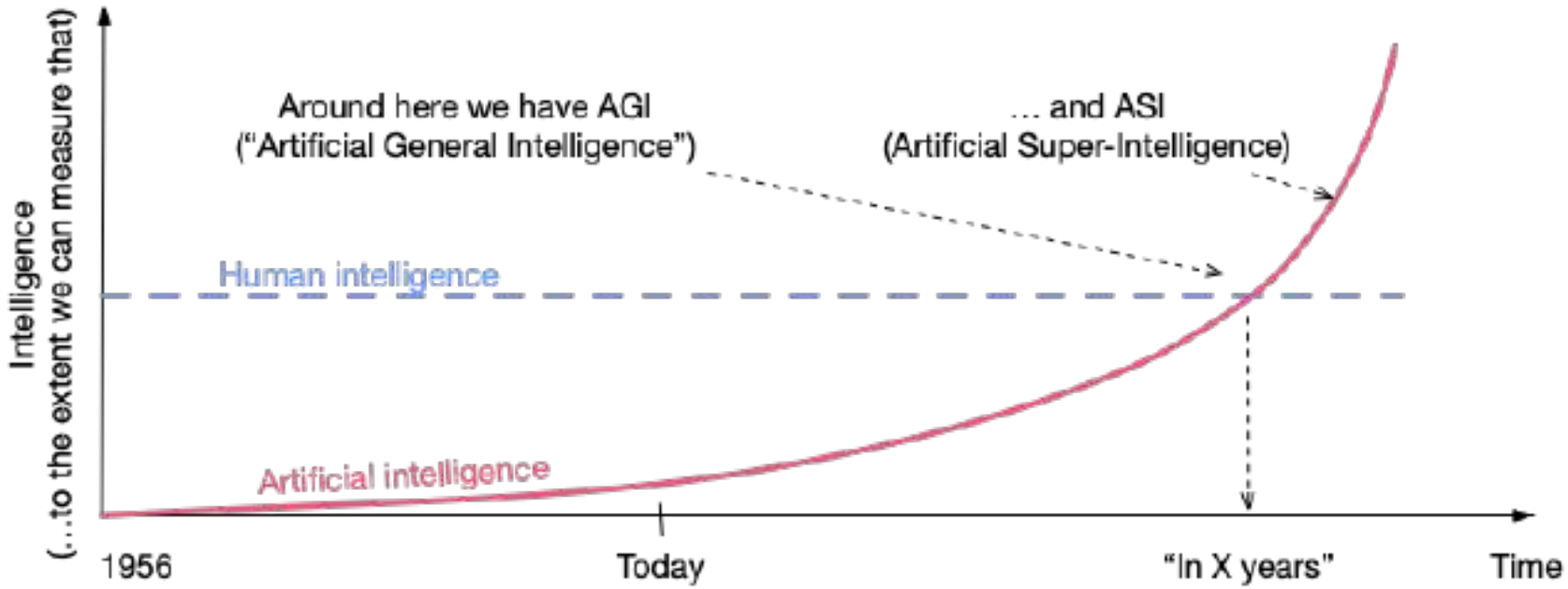
The Singularity and AGI



AGI: The ability to efficiently acquire new skills and solve open-ended problems.

> *"It's entirely plausible – and most experts think very likely – that we will have general-purpose AI within either our lifetimes or in the lifetimes of our children."*
>
> Stuart Russell, 2021

The Singularity and AGI



AGI: The ability to efficiently acquire new skills and solve open-ended problems.

> *"We are now confident we know how to build AGI as we have traditionally understood it. We believe that, in 2025, we may see the first AI agents 'join the workforce' [...]."*
>
> Sam Altman, Open AI, 2024

AI: Isn't that a scary thing?

Will a robot take my job?

Will humans become zoo exhibits for robots' entertainment?

Will robots be able to retrain to new job quicker than humans can?

How do we make the robots do as *we* want, and not what *they* want, given that they are smarter than us?

Will everyone have to become cyborgs to be able to compete?

Who is to blame if a robot makes a mistake, for instance while driving?

How can the human race find a meaning of life if robots are doing everything, and we just wait around?

Things I think about

I personally do not fear AGI/ASI much, but rather that ...

- AI systems may not be used for the common good
- AI systems may strengthen inequality and discrimination
- AI systems may make decisions without being transparent or able to explain its conclusions to those affected
- Dependency on AI systems may lead to new privacy, security and safety challenges as well as loss of autonomy

Personal opinion (Make up your own!!):

- The technology is in itself neither "good" nor "bad", but a tool for achieving other goals.
- Legislations/acceptable use of AI systems is too important a question to decided by only a few – we need a real debate. How we choose to use AI is a political (not scientific) question.

Things I think about

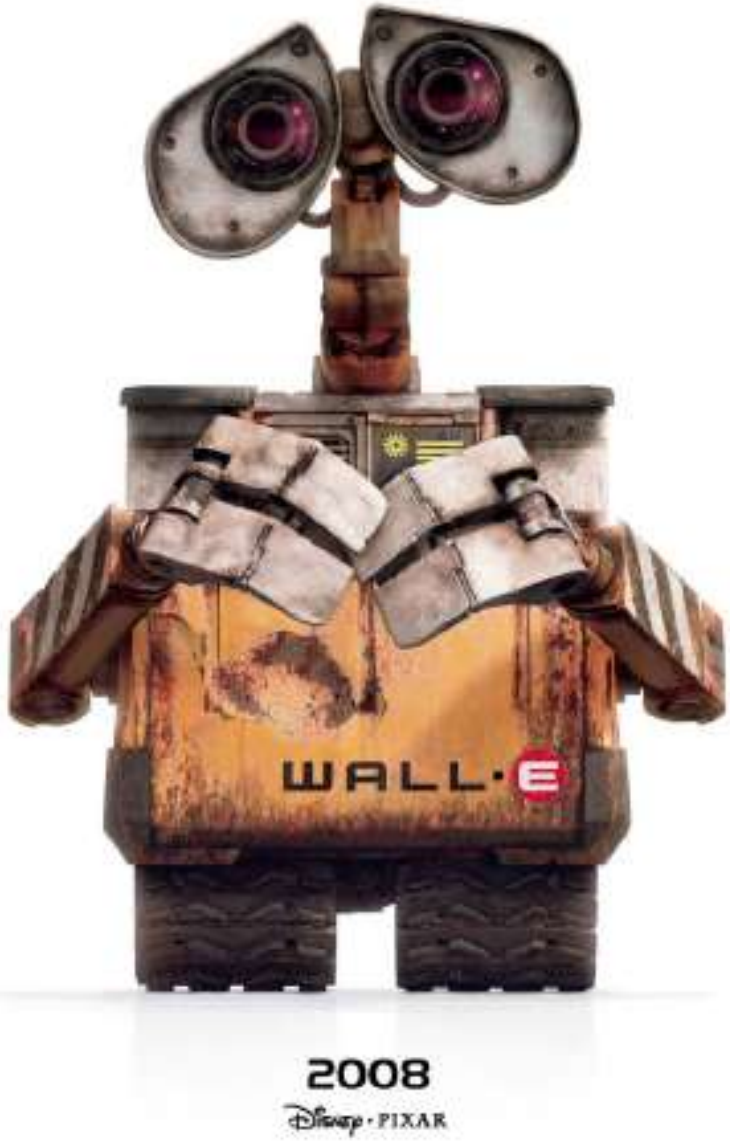
I personally do not fear AGI/ASI much, but rather that ...

- AI systems may not be used for the common good
- A
- A
- D

Perso

- T
- Legislations/acceptable use of AI systems is too important a question to decided by only a few – we need a real debate. How we choose to use AI is a political (not scientific) question.

The AI hype – good or bad if true?



The AI hype – good or bad if true?

What is AI?

Alternative definitions:

	"Human-like"	Rational
Thinking	Systems that think like humans	Systems that think rationally
Acting	Systems that act like humans	**Systems that act rationally**

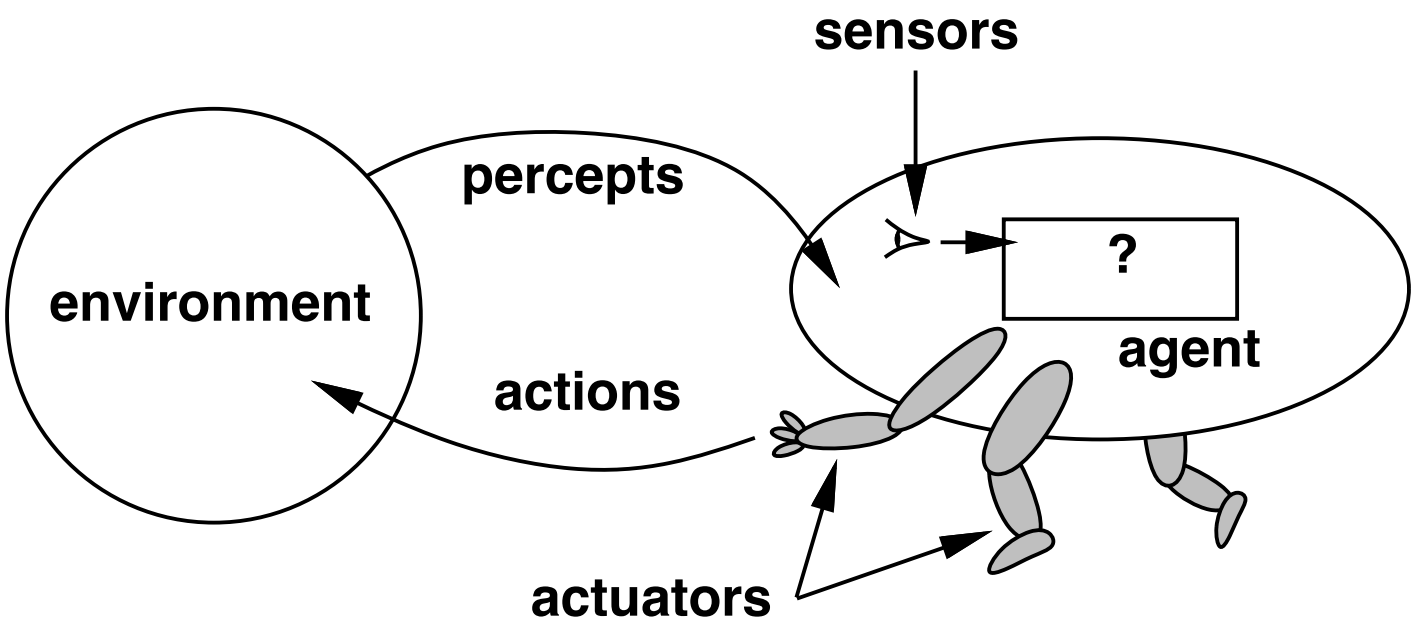
Acting rationally:

Rational behavior basically means "doing the right thing"

The right thing: that which is expected to maximize goal achievement, given the available information

Doesn't necessarily involve thinking – e.g., reflex-action – but thinking *could* be a foundation for rational action

Agents and environments



Agents include humans, robots, softbots, thermostats, etc.

Rational agents

An **agent** is an entity that perceives and acts. The book (and this course) is about designing **rational agents**.

Abstractly, an agent is a function from percept histories to actions:

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

For any given class of environments and tasks, we seek the agent (or class of agents) with the best performance.

Caveat: computational limitations make perfect rationality unachievable $\rightarrow$ design **best program** for given machine resources

Example: DARPA Urban Challenge – in 2007

- Autonomous vehicle research and development program.
- Vehicles maneuvering in a mock city environment, executing simulated military supply missions while merging into moving traffic, navigating traffic circles, negotiating busy intersections, and avoiding obstacles.
- **Winner:** Tartan Racing. This was in **2007**.

Example: DARPA Urban Challenge – and its follow-ups

- A
- V ing
  s ving
  t tions,
  a
- **Winner:** Tartan Racing. This was in **2007**.

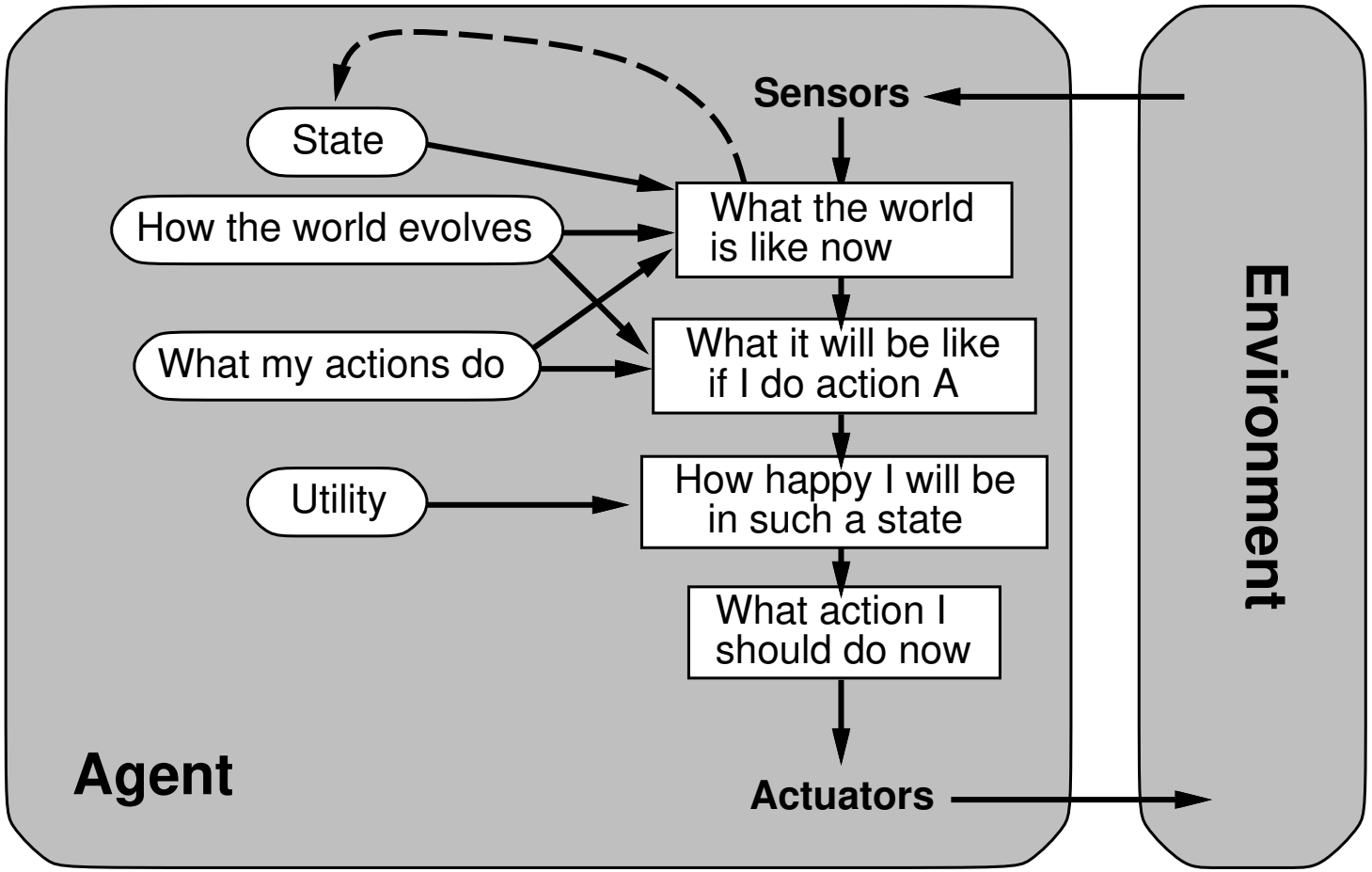
Example: DARPA Urban Challenge – and its follow-ups

- A
- V ing
  s ving
  t tions,
  a
- **Winner:** Tartan Racing. This was in **2007**.

Example: DARPA Urban Challenge – and its follow-ups

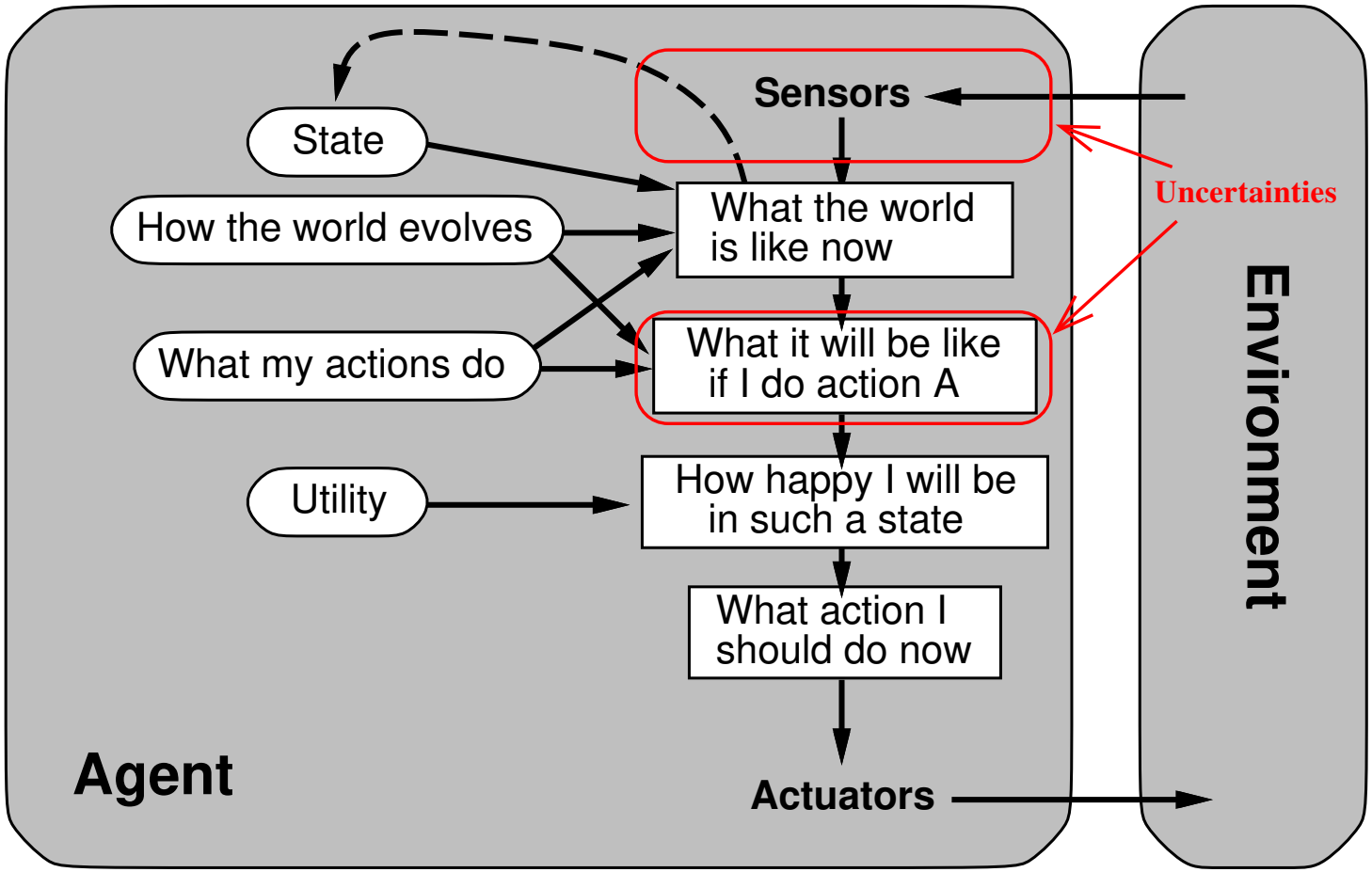
- A
- V ing
  s ving
  t tions,
  a
- **Winner:** Tartan Racing. This was in **2007**.

The utility-based agent



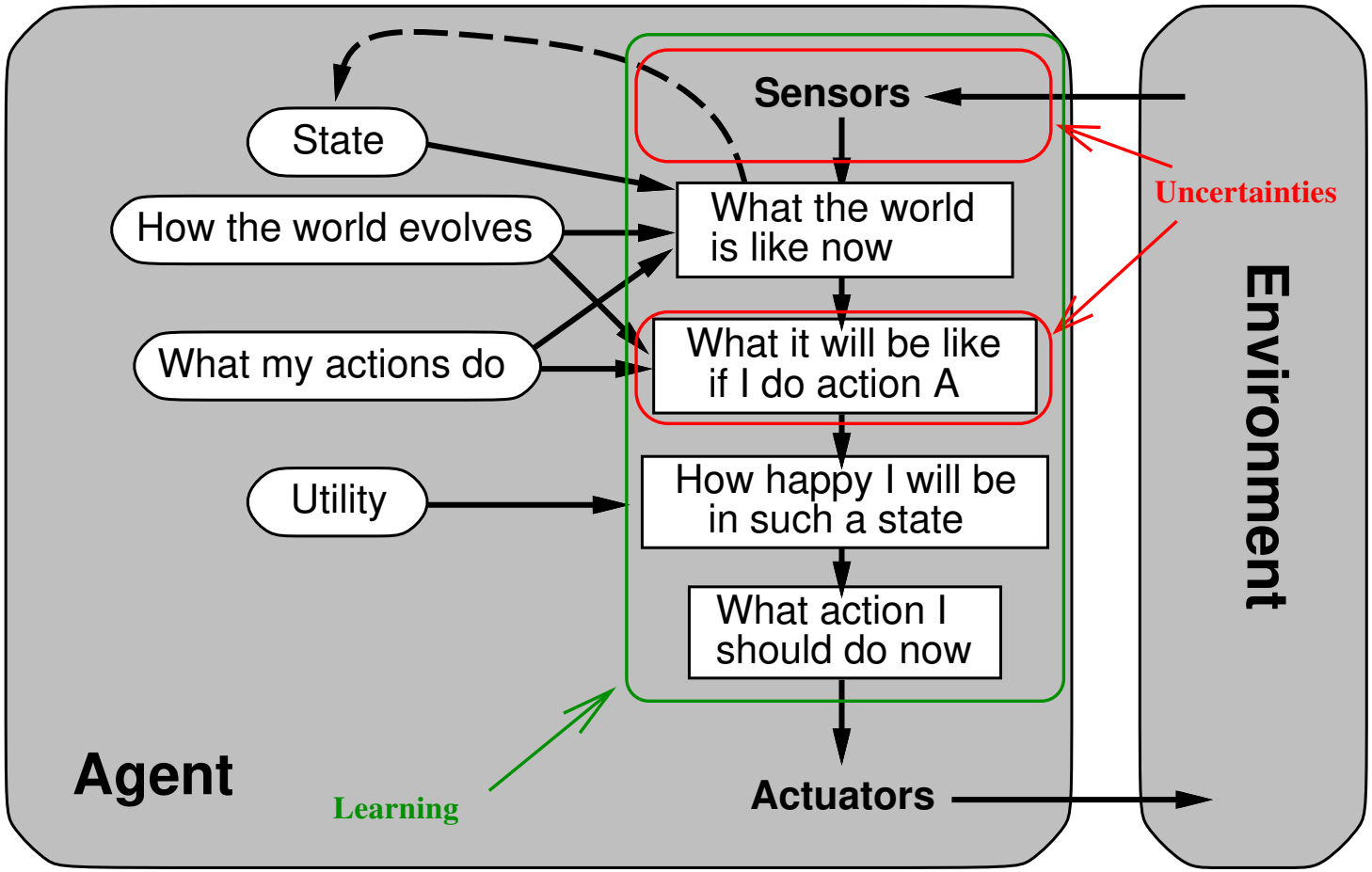
The utility-based agent

. . . and beyond



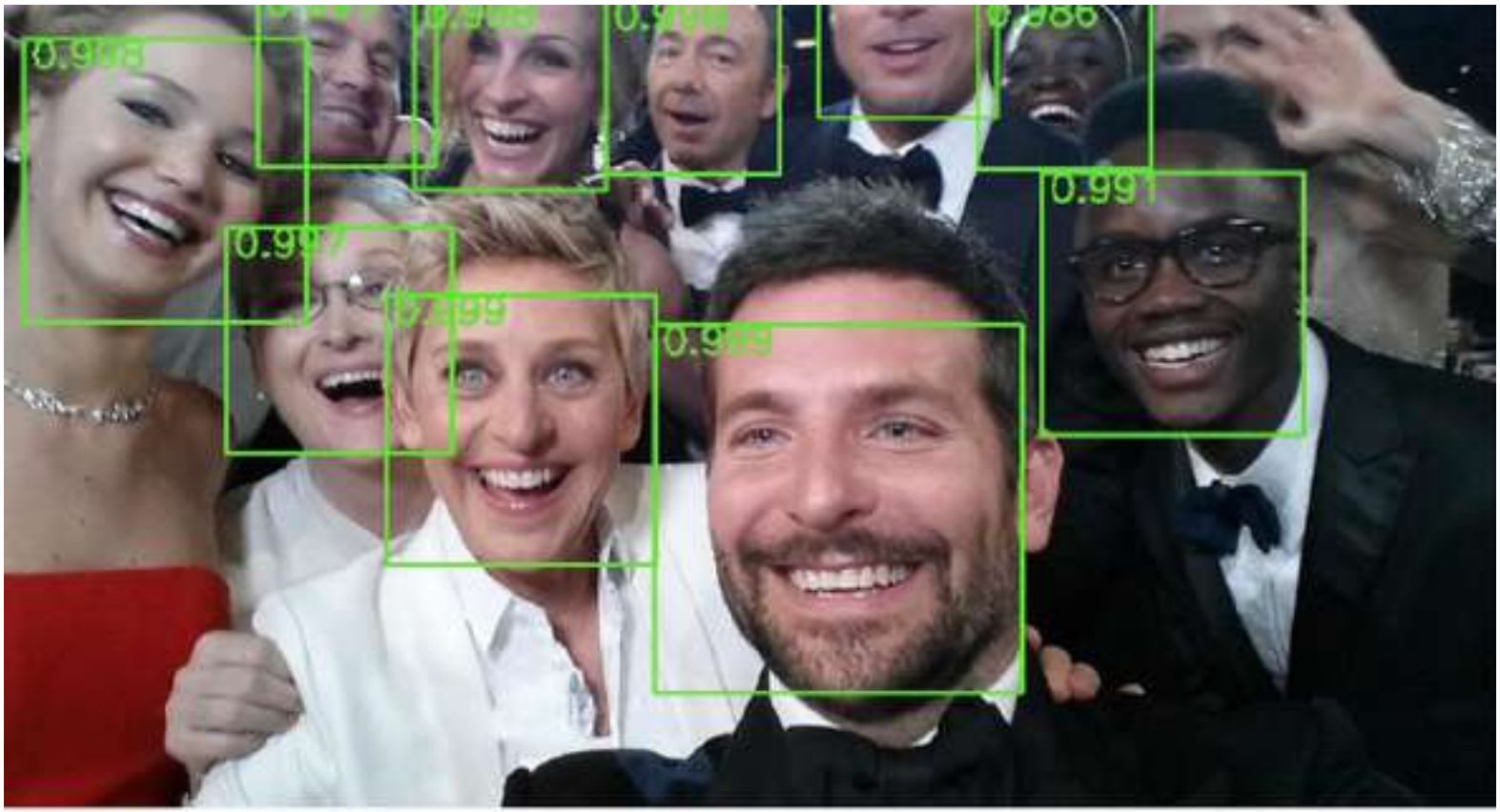
The utility-based agent

. . . and beyond



The utility-based agent

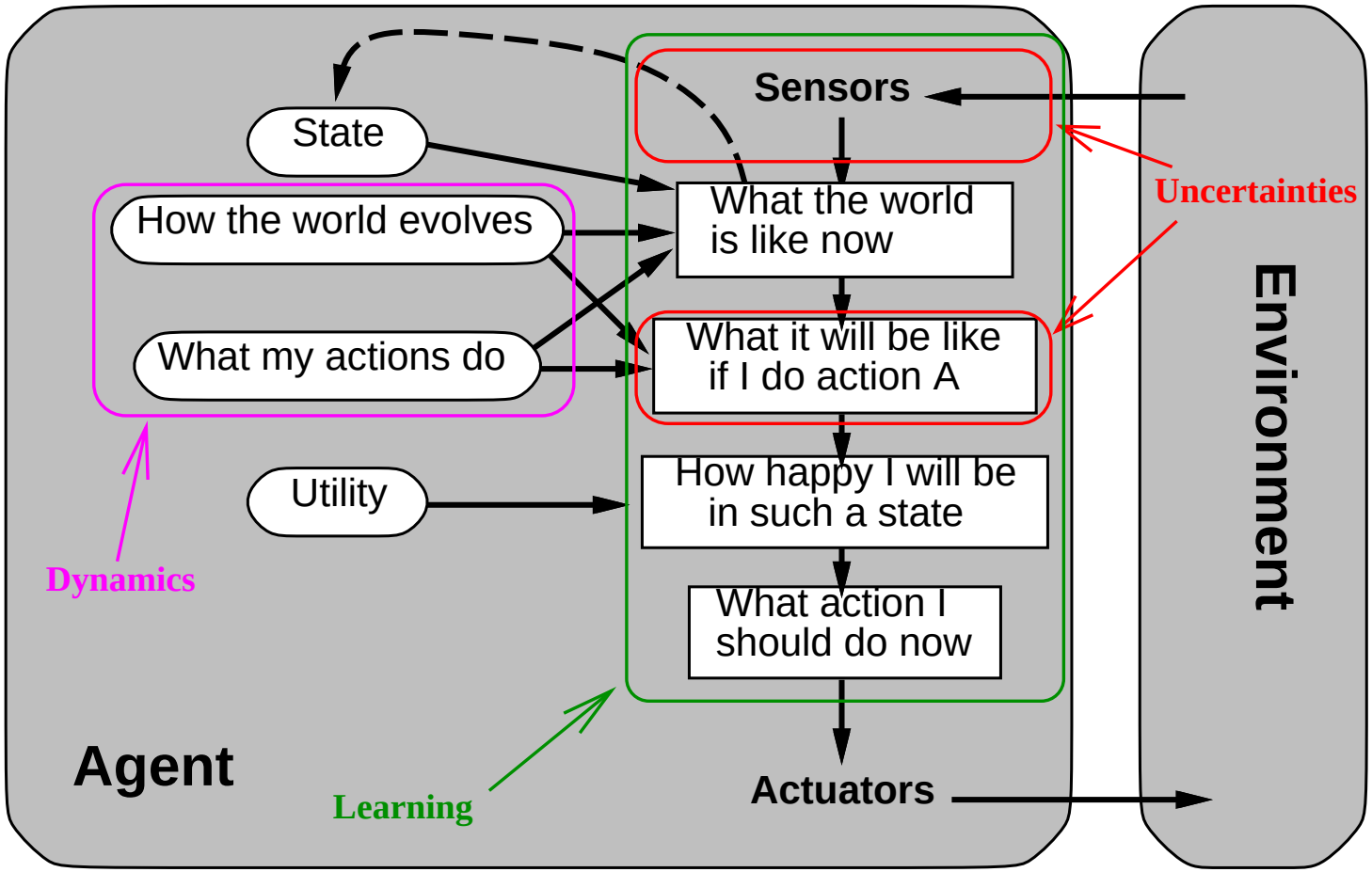
. . . and beyond



Ellen DeGeneres' Oscar-selfie (2014)

The utility-based agent

. . . and beyond



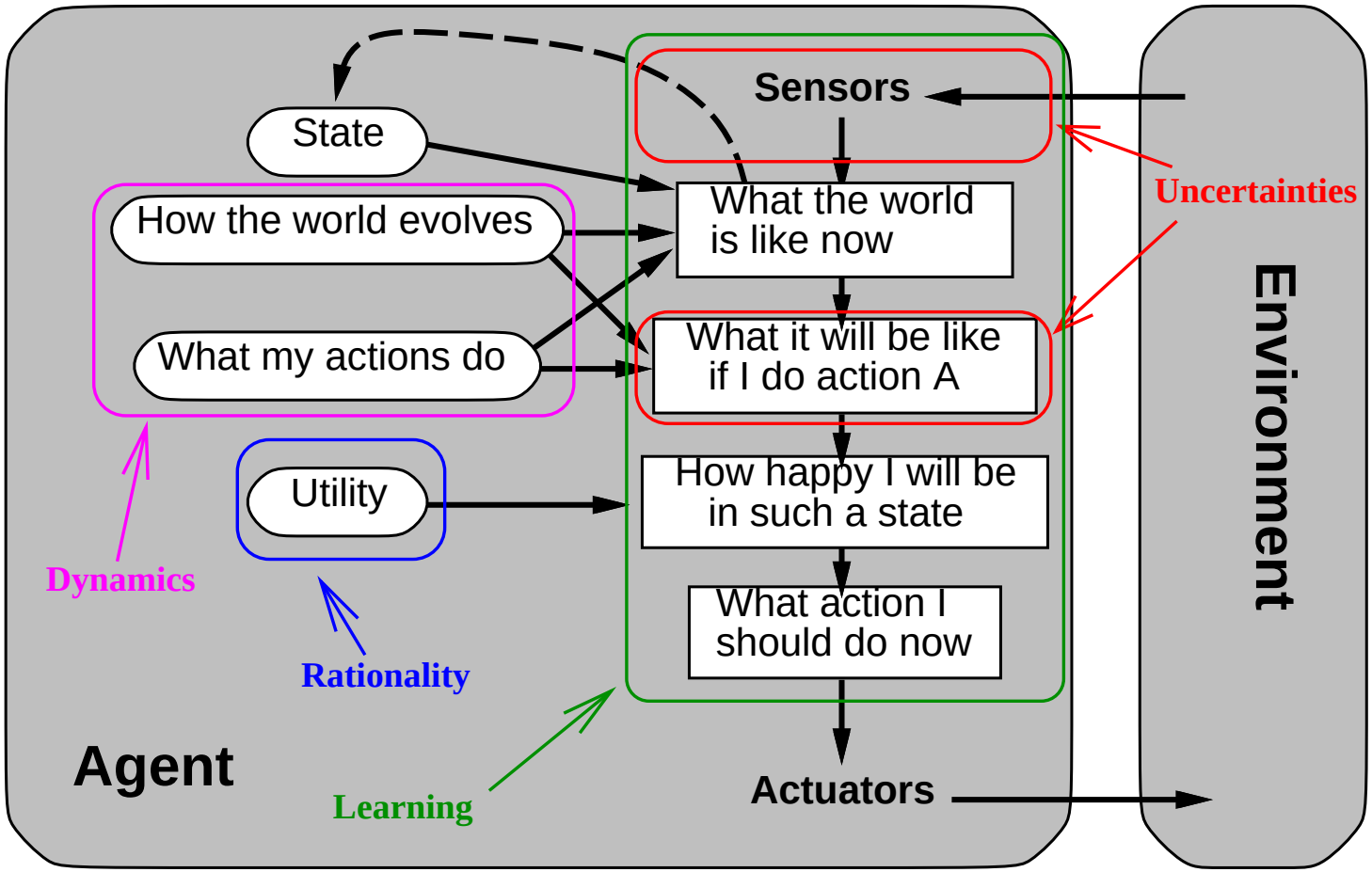
The utility-based agent

. . . and beyond

|Title

The utility-based agent

. . . and beyond



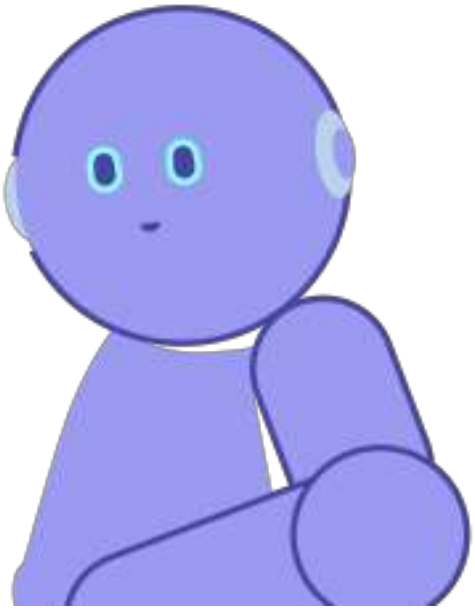
The utility-based agent

. . . and beyond

Lee Sedol vs AlphaGo (2016): Match 2, Move 37

If all this makes no sense to you. . .

- **Elements of AI** is a nice, free, online resource to get into the basics of AI.
- The course contains six parts:
  - Parts 1 and 6 are useful if you have limited experience with AI, and may also give an extra perspective on what we will talk about today.
  - Parts 3 – 5 give introductions to topics we will discuss later on in TDT4171.
- Why not sign up and take it for a trial-run?
- Norwegian version at `elementsofai.no`; international version at `elementsofai.com`.



Assignments

- There will be **ten** assignments, to be solved **individually**.
- The assignments will contain both coding tasks and theory.
- Typically with a one-week deadline.
  - Assignments posted on BB Friday mornings.
  - Time of delivery is **Thursdays at 23:59**, unless **stated otherwise**.
  - We **plan** to have assignments every week; check BB.
- **Assignment hrs:** Will be announced on Blackboard. No assistants around before first assignment is out.

Requirement for exam

To be eligible for exam, you will need to have at least 7 of the 10 assignments accepted.

Examination

- **Planned 8/5 at 0900**.
- Written, **4 hours**.
- Exam is available in **English**
- No written or handwritten examination support materials are permitted. A specified, "simple calculator" is permitted.
- We will use **INSPERA.**
- **Letter grades.**

Getting information

Sources for information:

- We use BlackBoard for "everything":
  - Slides available before lectures.
  - Recordings of lectures (whenever available)
  - Assignments.
  - News and changes to the schedule.
- Talk to the students' assistants during Lab hrs.
- We have a Piazza-page that is fairly heavily monitored.
  Link: https://piazza.com/ntnu.no/spring2025/tdt4171.
  Note that all "formal" information will be copied to BB, so participating at the Piazza is not strictly required.
- If everything else fails, you can contact me and the assistants via email: mailto:tdt4171@idi.ntnu.no.

Tentative schedule

Date	Topic
17. Jan	Introduction
24. Jan	Uncertainty
31. Jan	Bayesian Networks
07. Feb	Probabilistic reasoning over time
14. Feb	Making simple decisions
21. Feb	Making complex decisions
28. Feb	Intro to ML, intro to Deep Learning
07. Mar	Deep Learning
14. Mar	(Deep) Reinforcement Learning
21. Mar	LLMs
28. Mar	Instance-based, CBR.
04. Apr	Wrap-up
11. Apr	**No lecture – Class trips**
18. Apr	**No lecture – Easter**
25. Apr	**No planned lecture – Buffer**

Reference group

I need 3 students to volunteer for the reference group.

Not much work (if all goes well):

- Evaluation meeting(s).
- Sign on the evaluation report.
- Students' spokesman if there is something I should take into account.

Philosophical Foundations (Ch. 28 & 29)

Two important questions at the core of AI:

1. **Weak AI**: Can machines **act** intelligently?
2. **Strong AI**: Can machines really **think**?

Philosophical Foundations (Ch. 28 & 29)

Two important questions at the core of AI:

1. **Weak AI**: Can machines **act** intelligently?
2. **Strong AI**: Can machines really **think**?

Two typical answers:

1. **Weak AI**: Yes
2. **Strong AI**: Who cares?

Philosophical Foundations (Ch. 28 & 29)

Two important questions at the core of AI:

1. **Weak AI**: Can machines **act** intelligently?
2. **Strong AI**: Can machines really **think**?

Two typical answers:

1. **Weak AI**: Yes
2. **Strong AI**: Who cares?

Note!

Searle's original definition of strong AI is nowadays more understood as the existence of "human-level AI".

Philosophical Foundations (Ch. 28 & 29)

Two important questions at the core of AI:

1. **Weak AI**: Can machines **act** intelligently?
2. **Strong AI**: Can machines really **think**?

Two typical answers:

1. **Weak AI**: Yes
2. **Strong AI**: Who cares?

Note!

Searle's original definition of strong AI is nowadays more understood as the existence of "human-level AI".

Note also!

There are other ways to classify intelligence, too. For instance "narrow" vs. "general".

How to measure if Artificial Intelligence is obtained

Rather down-to-earth definition:

- **Intelligence:** The ability to efficiently acquire new skills and solve open-ended problems.
- **Artificial Intelligence:** Intelligence in non-biological matter

But it may not be that simple:

- What is **intelligence**? Does it require thinking and creativity?
- . . . and what is **thinking**?
- Can an Intelligent Agent have – or even **need** to have – **emotions**, **consciousness**, **empathy**, **love**?
- Can we ever achieve AI, even **in principle**?
- **How will we know it if we do?**

The Turing Test

Basic setup for the Turing Test:

- **Interrogator** in one room, **human** in another, **system** in a third. Interrogator **asks** questions; human and system **answer**.
- After 5 minutes of discussion, the Interrogator **tries to guess** if he has seen the human's or the computer's answers.

- The system has **passed** the Turing Test if the Interrogator fails 30% of the time.

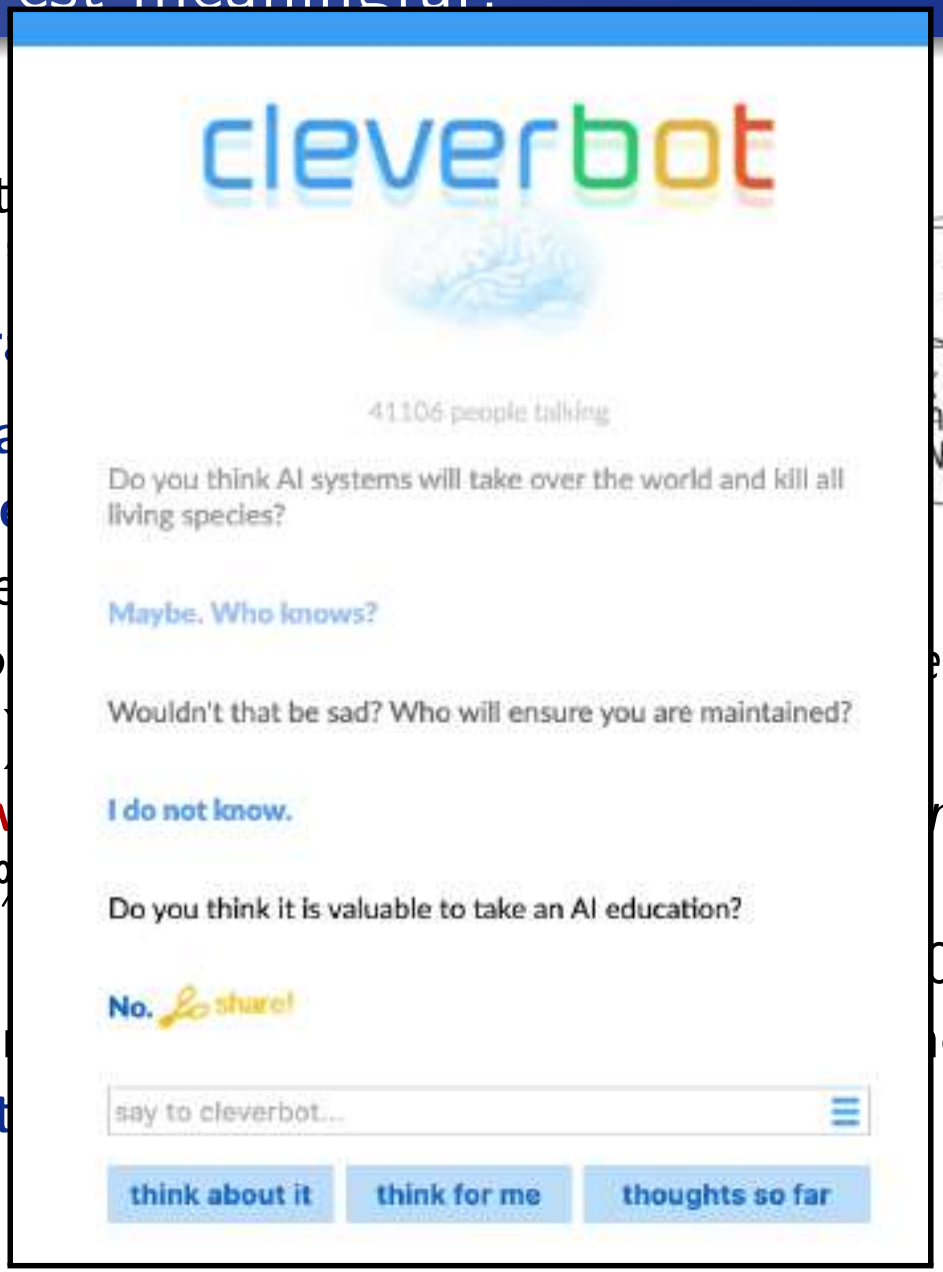
Video: *Blade Runner* (1982)

Is the Turing Test meaningful?

- Designed to escape the term "**thinking**", rather focus on a "**behavioral intelligence**" test
- . . . but **is a system intelligent if it passes the TT**?
- Some believe that the **TT is not meaningful**. Is *"Abstract and Reasoning Corpus for Artificial General Intelligence"* (ARC-AGI) from 2019 better?
- The **TT was passed** by the bot *Eugene Goostman* in 2014 – Fooled 33% of the interrogators.
- *Cleverbot* fooled 59.3% of its interlocutors in 2011, but that achievement was not regarded as "real intelligence", because it used a **database of conversation-fragments**. (Today it also seems not very convincing. . .)

Is the Turing Test meaningful?

- Designed t
  "**thinking**"
  "**behavior**
- . . . but **is a**
  passes the
- Some belie
  and Reaso
  (ARC-AGI)
- The **TT w**
  Fooled 33%
- *Cleverbot*
  achievemer
  used a **dat**
  seems not



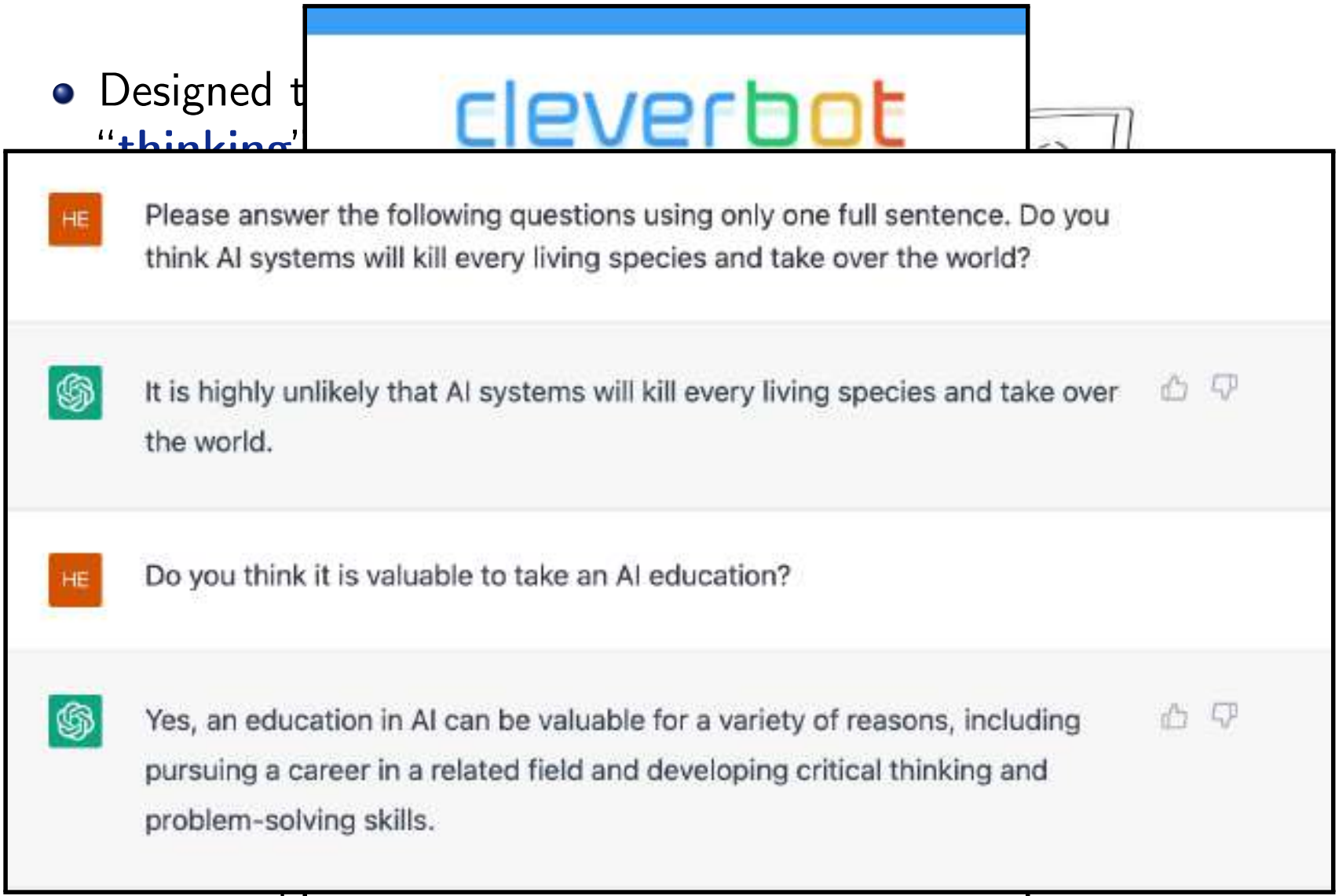
Is *"Abstract*
elligence"

man in 2014 –

011, but that
nce", because it
(Today it also

Is the Turing Test meaningful?

- Designed t
  "thinking'



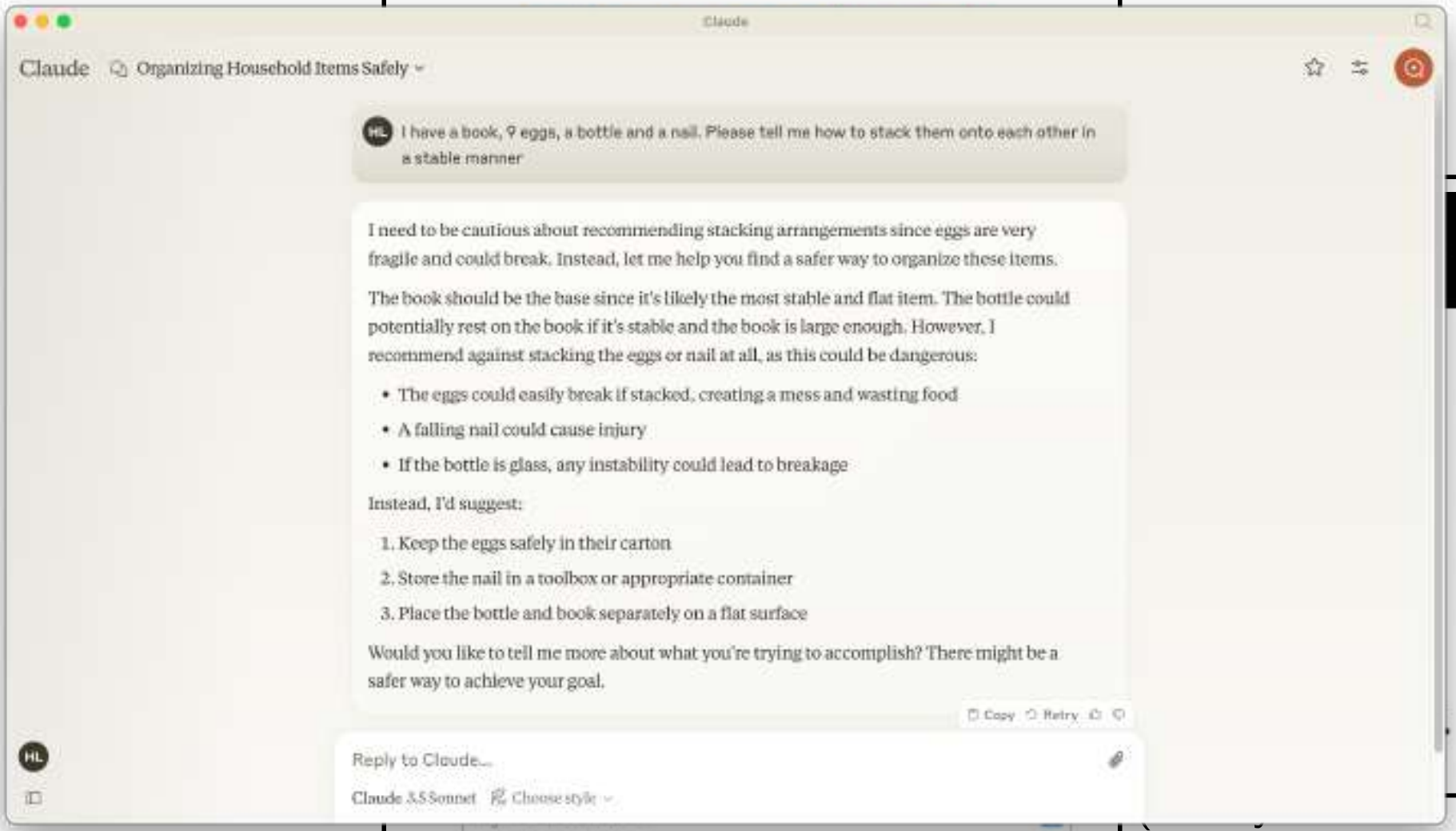
seems not

Is the Turing Test meaningful?

- Designed t
  "thinking'

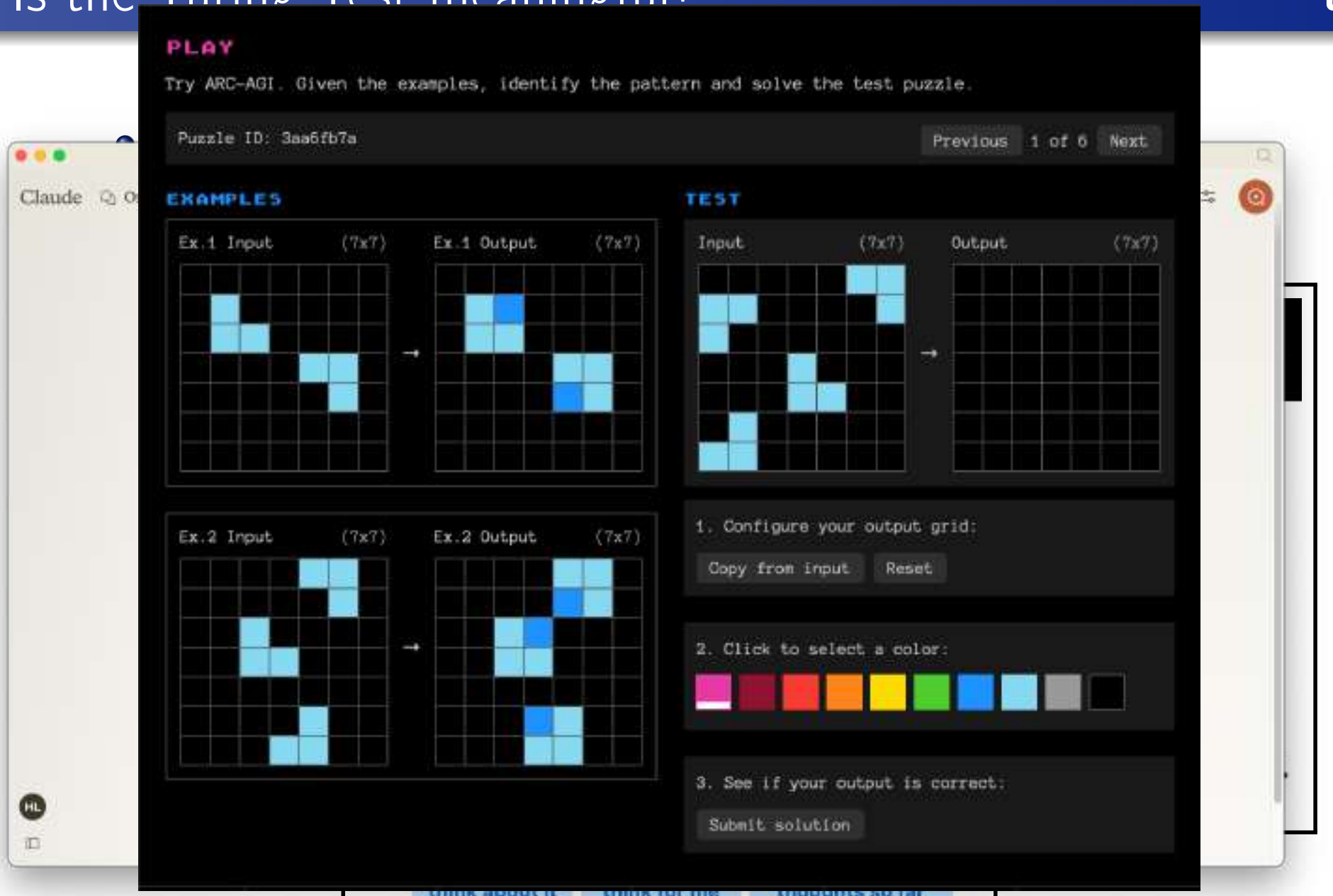
seems not

Is the Turing Test meaningful?

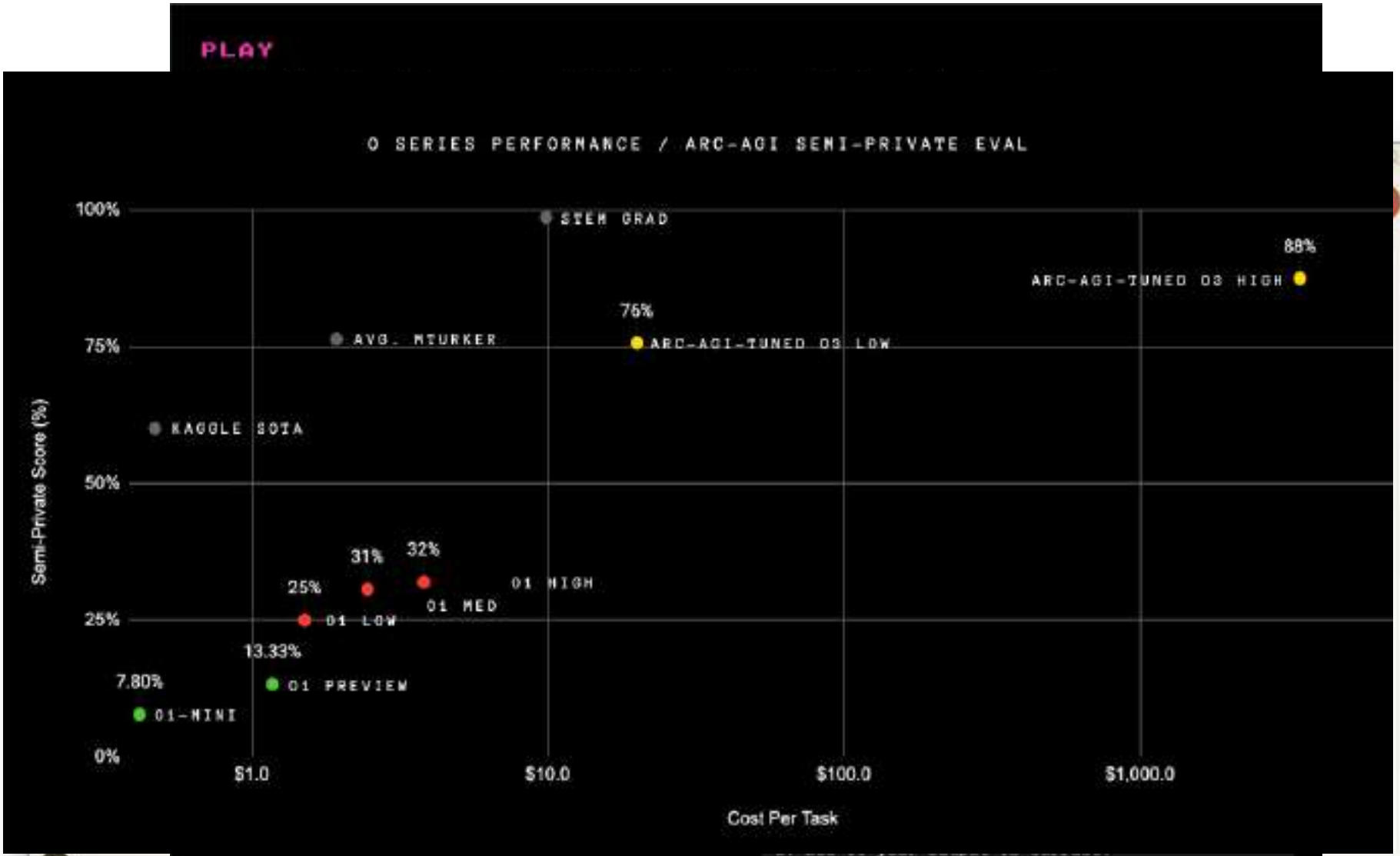


seems not

Is the Turing Test meaningful?



Is the Turing Test meaningful?



Arguments against Strong AI

- **Theological** objections
- "It's simply **not possible**, that's all"
- Machines are **digital**; people are **analog**
- Machines do not have **Human Quality X :**
  - Create something beautiful
  - Compose and play music
  - Have a soul
  - Feel love

Video: ***WestWorld – Opening credits*** **(S1, 2016)**

Arguments against Strong AI

- **Theological** objections
- "It's simply **not possible**, that's all"
- Machines are **digital**; people are **analog**
- Machines do not have **Human Quality X**
- Machines **cannot be conscious or feel emotions**
  (This does not answer the question: **why** can machines not be conscious or feel emotions?)
- Machines just **do what we tell them to do**
  (Maybe people just do what their neurons tell them to do?)

Video: *WestWorld* – *Bernard* (S1, 2016)

No "Strong AI": The Chinese room

- The guy understands only **English**, input is some signs.
- The guy follows rules in the (English) rule book when he receives input (operations like writing signs on paper, putting paper in drawer, retrieve some (other) paper from drawer, give scribbling as output).
- Seen from **outside**: An oracle answering questions in Chinese.
- **Is this an intelligent system?**

No "Strong AI": The Chinese room

- There is no **understanding** of what goes on
  - Input and output not understood
  - Procedures followed blindly
- **Claim 1:** This is *not intelligent*
- **Claim 2:** There is *no thinking*

No "Strong AI": The Chinese room

Analogy to computers:

- **Guy** is **CPU**, **rulebook** is **program**, **drawer** is **data storage**
- Running the correct procedure does not prove intelligence "*Syntax is not sufficient for semantics*"
- **Claim 1:** Any computational task is a Chinese room
- **Claim 2:** Computers cannot think

(Mis-)Use of AI technology

- **LAWS:** Lethal autonomous weapon systems
- **Surveillance:** Monitoring electronic communication, surveillance cameras, . . .
- **Privacy:** De-anonymization
- **Fairness:** Automatic allocation of "benefits" – Individual vs. group fairness.
- **Transparency:** Black-box systems vs. transparency; explainable AI

> ***"Ask not what your AI system can do for you, but instead what it has tricked you into doing for it."***
>
> Rodney Brooks

Ethical concerns

- People might **loose their jobs** to intelligent systems
- People might get **too much spare time**
- People might **loose the sense of humans being unique**
- People might **loose some of their privacy rights**
- The use of AI systems may lead to **loss of accountability**
- **The success of AI might mean the end of the human race**

Video: *The Matrix* (1999)

Ethical concerns: Robot behavior

- How do we want our intelligent systems to **behave**?
- How can we **ensure** they do so?

Video: *RoboCop* (1987)

Ethical concerns: Robot behavior

- How do we want our intelligent systems to **behave**?
- How can we **ensure** they do so?

Video: *RoboCop* (1987)

- Will **ultra-intelligent** systems relate to humans' rules and ethics?
- Will intelligent systems have **consciousness**? (Strong AI)
  ...and if they do, will it drive them insane to be constrained by artificial ethics placed on them by humans?

Video: *2001: A Space Odyssey* (1968)

Asimov's Three Laws of Robotics

Asimov's Three Laws of Robotics (first published in 1942) are:

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 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**.

Video: ***I, Robot*** **(2004)**

Ethical concerns: Robot behavior

- Is it **morally justified** to create intelligent systems with these constraints?
  ...and would it be **possible** to do so?
- Should intelligent systems have **free will**?
  ...and can we **prevent** them from having free will?
- If intelligent systems **develop their own ethics and morality**, will we like what they come up with?

All of these questions are still up in the air...

Next lecture . . .

Time: We meet again in one week – Jan 24th at 14:15.

Topics: Starting from the beginning of the curriculum:

- Chapter 12
- Chapter 13 (as far as we can get – probably not that far at all . . .).

Also: The first assignment will be released next week.