University of South-Eastern Norway
School of Business

# Lecture 10: Philosophy, ethics, and safety of AI

Sinuo Wu

Course: AI for Business Applications (AI3000)

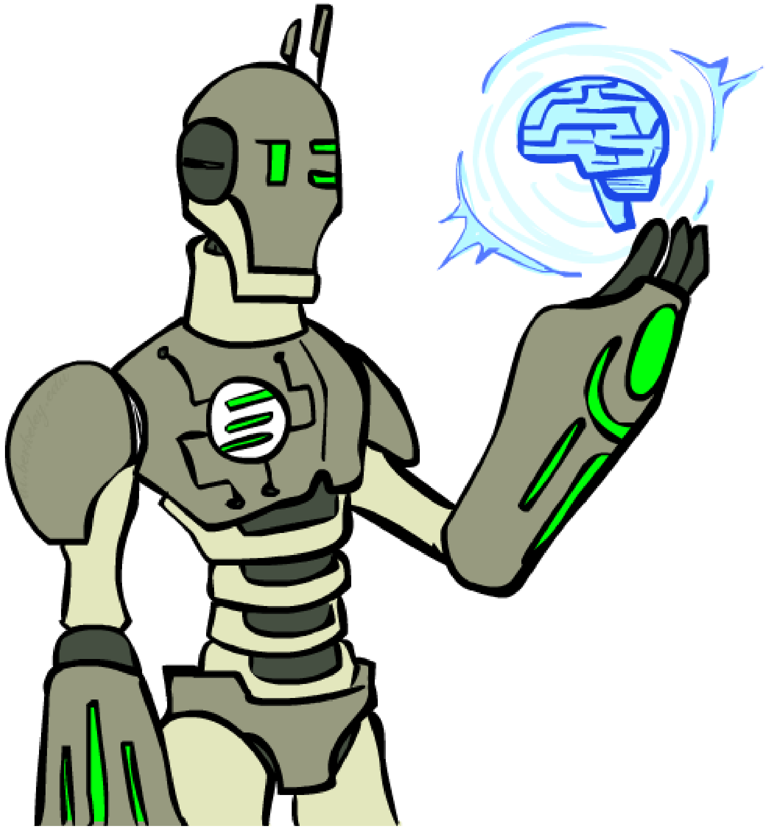
# Quiz

- Enter room number or Scan the QR code

# In this lecture

- *In which we consider the big questions around the meaning of AI, how we can ethically develop and apply it, and how we can keep it safe*

- The Limits of AI
- Can AI really think?
- The Ethics of AI



# The limits of AI

# What AI can and cannot do

01

## What AI can do

- Generate quick results
- Scan large databases and search facts in few seconds
- Find mathematical and logical solutions with few errors
- Make decisions based on solely objective criteria

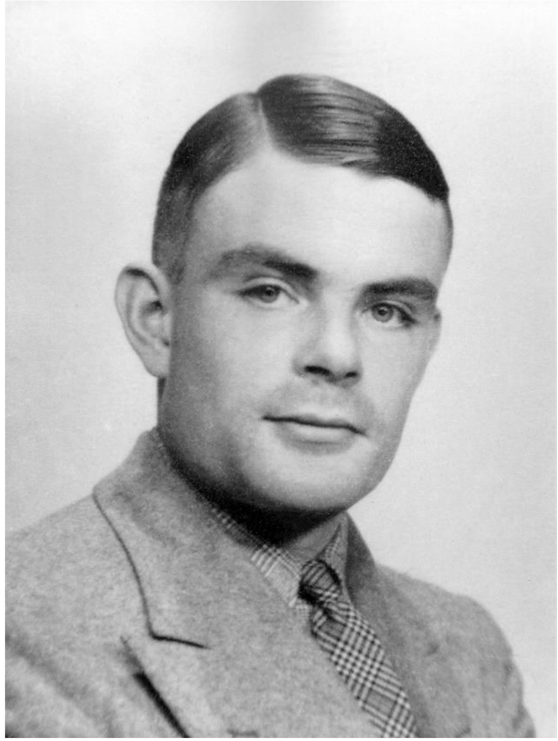
02

## What AI cannot do

- Generate results with common-sense reasoning
- Develop own unique insights
- It lacks moral and ethical judgment.
- Engage in social interactions and communications with human emotions.
- AI lacks self-awareness and consciousness.
- AI lacks genuine creativity and innovation.

Everything is by computing!

# The Limits of AI



Alan Turing (1912-1954)

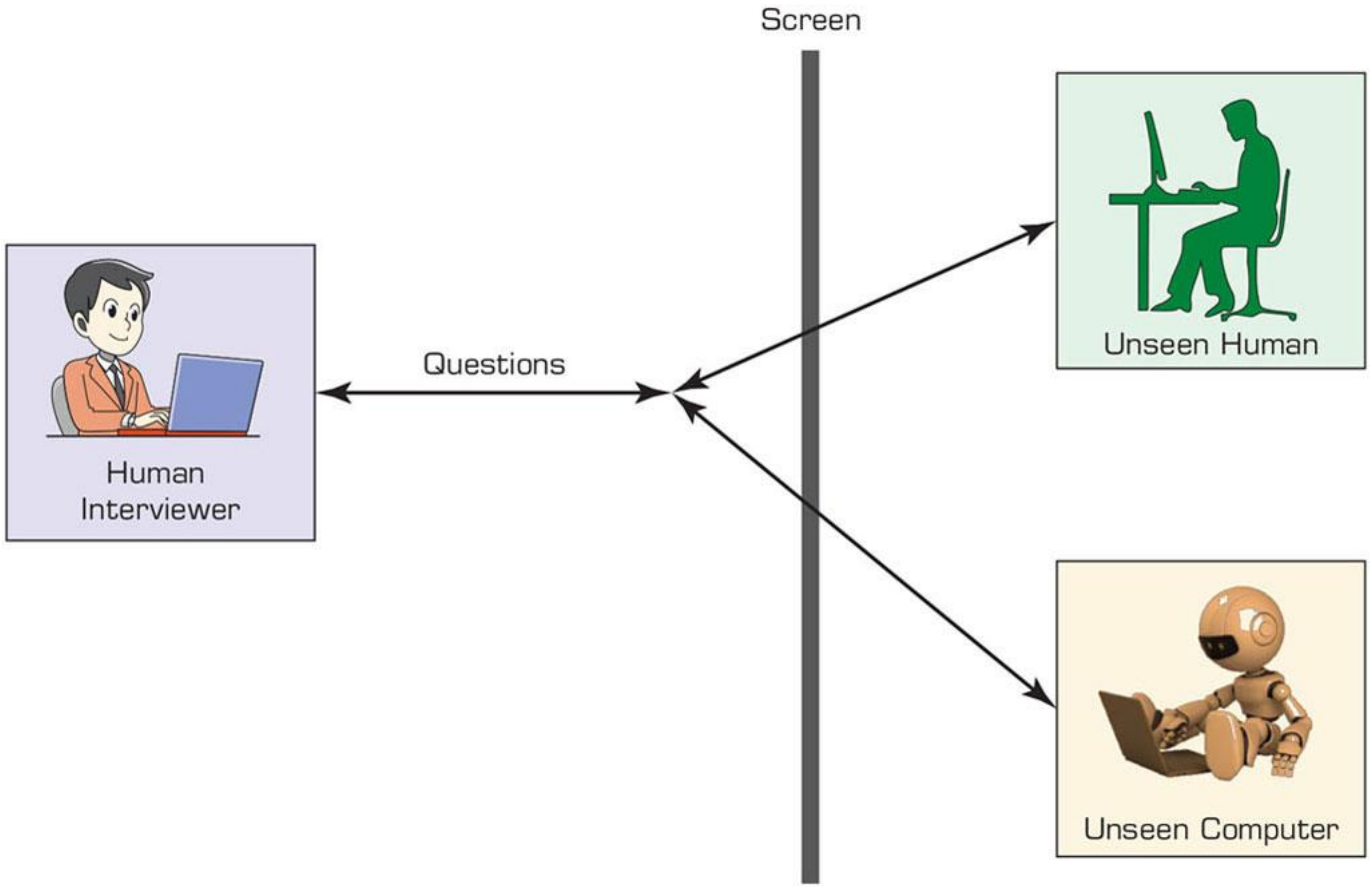
Turing was the first to raise possible objections to AI

**Weak AI**—the idea that machines could act as if they were intelligent
**Strong AI**—the assertion that machines that do so are actually consciously thinking (not just simulating thinking).

The argument from **informality**:
Human behavior is far too complex to be captured by any formal set of rules.

# Human and Computer Intelligence

- Measuring AI



# The Limits of AI

- The argument from **disability:**
  - “a machine can never do X”

As examples of X, Turing lists the following:

> Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as man, do something really new

# The Limits of AI

Kenneth Sayre (1928-2022)

"Artificial intelligence pursued within the cult of computation stands not even a ghost of a chance of producing durable results." (1993)

- Good Old-Fashioned AI (GOFAI)
  - Programming AI with logical rules
  - Logical rules can never track emotional human

# The Limits of AI

Hubert Dreyfus (1929-2017)

- What computers can’t do (1972)
- What computers still can’t do (1992)
- Mind over machine (1986)

Ex. “Dog(x) ⇒ Mammal(x)”

If x is a dog, then x is a mammal.

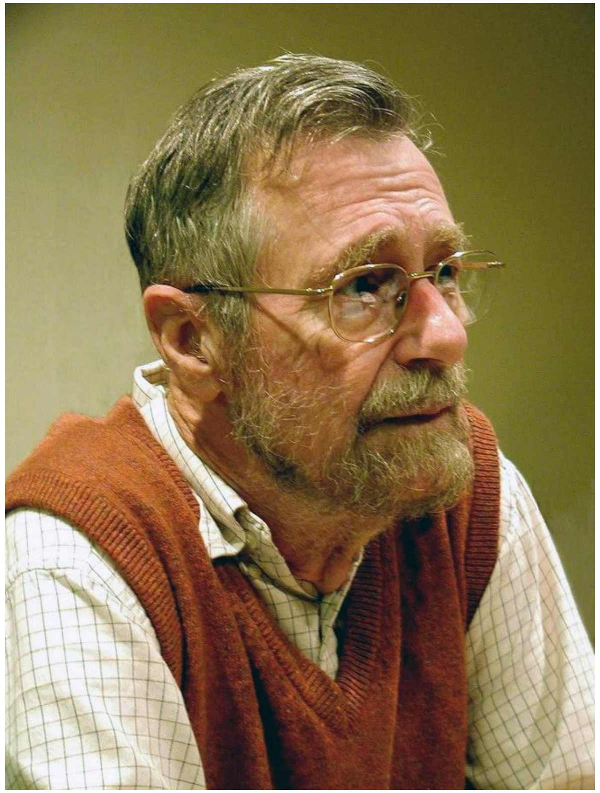
CNN model?

# Can AI Really Think

# Can AI really think

- **Consciousness**
  - Consciousness involves an awareness of both external stimuli (the world around us) and internal mental states (our thoughts, emotions, and self-awareness).
  - "Consciousness is the greatest movie-maker in the whole world." - Amit Ray
- **Qualia**
  - The subjective qualities of our conscious experiences, such as the taste of chocolate, the feeling of warmth, or the color red. They are difficult to describe to others because they are purely subjective.

# Can AI really think?



Edsger Dijkstra (1984)

Some philosophers claim that a machine that acts intelligently would not be *actually* thinking. But it would be only a *simulation* of thinking.

- Can submarines swim?
- Can airplanes fly?

# Can AI really think

**The Translation Room**

- A human, who understands only English, inside a room that contains a rule book, written in English, and various stacks of paper
- Pieces of paper containing symbols are slipped under the door to the room
- The human follows the instructions in the rule book, finding symbols in the stacks, writing symbols on new pieces of paper, rearranging the stacks, and so on
- Passed back to the outside world
- It is given that the human does not understand Chinese
- Computer are in essence doing the same thing, so therefore computers generate no understanding.

https://course.elementsofai.com/1/3

# The Ethics of AI

# The Ethics of AI

- Given that AI is a powerful technology, we have a moral obligation to use it well, to promote the positive aspects and avoid or mitigate the negative ones

- **Positive aspects examples**
  - AI can save lives through improved medical diagnosis, new medical discoveries, better prediction of extreme weather events

  - AI can improve lives. Microsoft's AI for Humanitarian Action program applies AI to recovering from natural disaster

  - AI applications in crop management and food production help feed the world

# The Ethics of AI

- **Negative aspects example**
- **Lethal autonomous weapons**

  - The UN defines a lethal autonomous weapon as one that locates, selects, and engages (i.e., kills) human targets without human supervision.
  - Autonomous weapons have been called the “third revolution in warfare” after gunpowder and nuclear weapons. Their military potential is obvious
  - The debate over autonomous weapons includes **legal, ethical and practical aspects.**

# Surveillance, security, and privacy

- As of 2018, there were as many as 350 million surveillance cameras in China and 70 million in the United States.

- As more of our institutions operate online, more vulnerable to cybercrime. Attackers can use automation to probe for insecurities and they can apply reinforcement learning for phishing attempts and automated blackmail

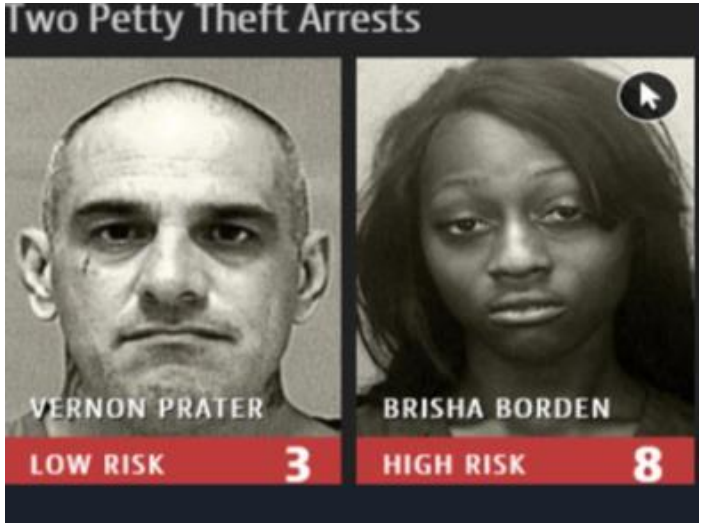
- Defenders can use unsupervised learning to detect anomalous incoming traffic patterns and various machine learning techniques to detect fraud

# Fairness and bias

- Machine learning models can have societal bias
- Designers of machine learning systems have a moral responsibility to ensure that their systems are fair
- Six of the most commonly-used concepts for fairness:
  - Individual fairness
  - Group fairness
  - Fairness through unawareness
  - Equal outcome
  - Equal opportunity
  - Equal Impact

# Training Bias

- Bias come from Original Datasets or from Real World interactions.
- They can impact your real life.



- Datasets must be revisited before each retraining.

# Trust and transparency

- People need to be able to trust the systems they use

- Engineered systems must go through a verification and validation (V&V) process
  - Verification means that the product satisfies the specifications
  - Validation means ensuring that the specifications actually meet the needs of the user and other affected parties

- Certification and safe standards, ISO in other industries

- The AI industry is not yet at this level of clarity, although there are some frameworks in progress, such as IEEE P7001, a standard defining ethical design for artificial intelligence and autonomous systems

- **Transparency**: consumers want to know what is going on inside a system, and that the system is not working against them.

# Trust and transparency

An AI system that can explain itself is called explainable AI (XAI).

A good explanation has several properties:

- it should be understandable and convincing to the user.
- it should accurately reflect the reasoning of the system.
- it should be complete.
- it should be specific in that different users with different conditions or different outcomes should get different explanations.

# Robot rights

- if robots can feel pain, if they can dread death, if they are considered "persons," then the argument can be made that they have rights and deserve to have their rights recognized
- If robots have rights, then they should not be enslaved, and there is a question of whether reprogramming them would be a kind of killing
- Another ethical issue involves voting rights: a rich person could buy thousands of robots and program them to cast thousands of votes—should those votes count?
- Ernie Davis argues for avoiding the dilemmas of robot consciousness by never building robots that could possibly be considered conscious

# AI Safety

- Almost any technology has the potential to cause harm in the wrong hands, but with AI and robotics, the hands might be operating on their own.

# AI in Business

# Case 1

**Case Overview:**

- Amazon developed an AI tool to automate its hiring process.
- The AI system was biased against women because it was trained on a decade of resumes, primarily from men.
- The bias occurred as the system prefers male candidates due to historical hiring patterns.

**Consequences:**

- Unintended reinforcement of gender bias in hiring.
- Ethical and practical concerns about gender disparities.
- Amazon abandoned the AI tool for hiring.

**Lessons Learned:**

- The importance of rigorous testing and oversight in AI development.
- The need to curate training data to prevent bias.
- The critical role of diversity in tech to avoid discrimination.

# Case 2

**Case Overview:**

- Facebook's News Feed algorithm determines the content shown to users.
- It uses complex algorithms to prioritize and personalize content based on user behaviour.

**Consequences:**

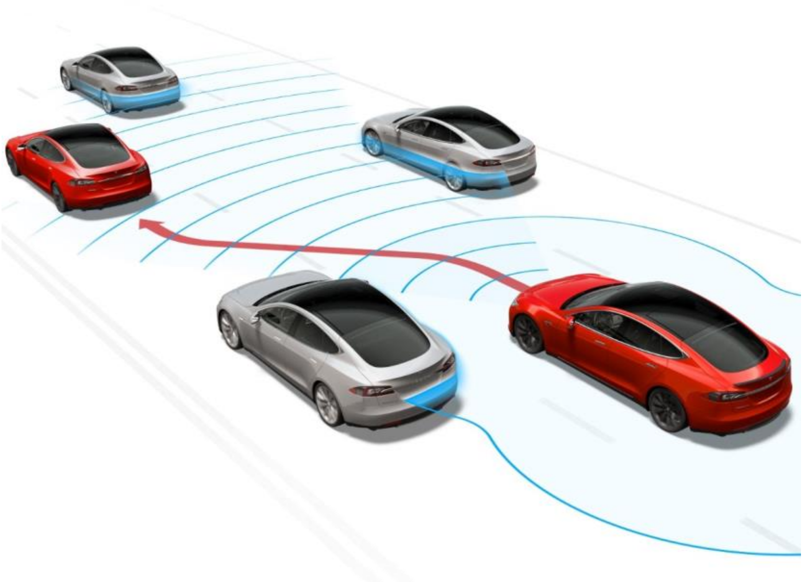
- This algorithm can create "filter bubbles," reinforcing users' existing beliefs and preferences.
- Concerns about the balance between user engagement and responsible content.

**Lessons Learned:**

- The importance of transparency in how algorithms prioritize content.
- The ethical responsibility in shaping the information users encounter on social media.

Facebook ads
...vertising on Facebook makes it easy to find the right people,
capture their attention and get results.
Create an Ad

# Case 3



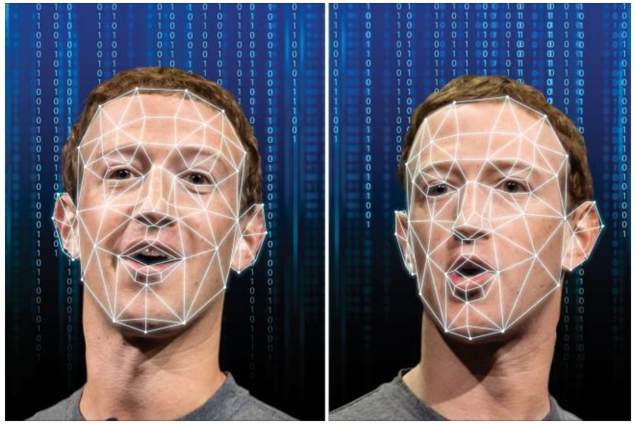
**Case Overview:**

- Tesla's Autopilot system controls self-driving cars.
- It aims to enhance convenience and safety for drivers.

- **Consequences:**
- Accidents involving Tesla's Autopilot system have raised questions about safety and liability.
- Ethical and legal debates about the role of human oversight in autonomous driving technology.

- **Lessons Learned:**
- The need for continuous safety monitoring and improvement in autonomous vehicle technology.
- Balancing innovation with the ethical responsibility for safety and accountability.

# Case 4



**Case Overview:**

- Deepfake technology can create highly convincing but fabricated content.
- It is increasingly used in marketing and advertising to engage consumers.

**Consequences:**

- Ethical dilemmas surrounding the deceptive use of deepfake technology in marketing.
- Concerns about transparency, trust, and responsible content creation.
- Concerns about job replacement in entertainment industry. (eg. Recent strike from actors to resist AI)

**Lessons Learned:**

- The importance of clear guidelines and ethical standards in content creation and advertising.
- The need for transparency and disclosure when using deepfake technology for marketing purposes.

# Case 5

**Case Overview:**

- AI systems are used in healthcare for medical diagnoses and treatment planning.
- They analyze medical data, images, and patient information.

**Consequences:**

- Ethical concerns related to patient data privacy and security.
- Addressing biases in AI algorithms and maintaining human oversight in healthcare decisions.

**Lessons Learned:**

- Prioritizing patient privacy, data security, and informed consent.
- Ensuring that AI in healthcare augments the expertise of medical professionals and doesn't replace it entirely.

# The future of work

- An immediate reduction in employment when an employer finds a mechanical method to perform work previously done by a person
- More automation with physical robots, first in controlled warehouse environments, then in more uncertain environments, building to a significant portion of the marketplace by around 2030.
- The ratio between workers and retirees' changes. In 2015 there were less than 30 retirees per 100 workers; by 2050 there may be over 60 per 100 workers
- Problems due to the pace of change
- Rules and Regulations for fast changing AI

# Summary

- Philosophers use the term weak AI for the hypothesis that machines could possibly behave intelligently, and strong AI for the hypothesis that such machines would count as having actual minds.

- AI is a powerful technology, and as such it poses potential dangers, through lethal autonomous weapons, security and privacy breaches, unintended side effects, unintentional errors, and malignant misuse. Those who work with AI technology have an ethical imperative to responsibly reduce those dangers.

- AI systems must be able to demonstrate they are fair, trustworthy, and transparent.

- There are multiple aspects of fairness, and it is impossible to maximize all of them at once. So, a first step is to decide what counts as fair

- Automation is already changing the way people work. As a society, we will have to deal with these changes.

Have a Break!

# Inspiration