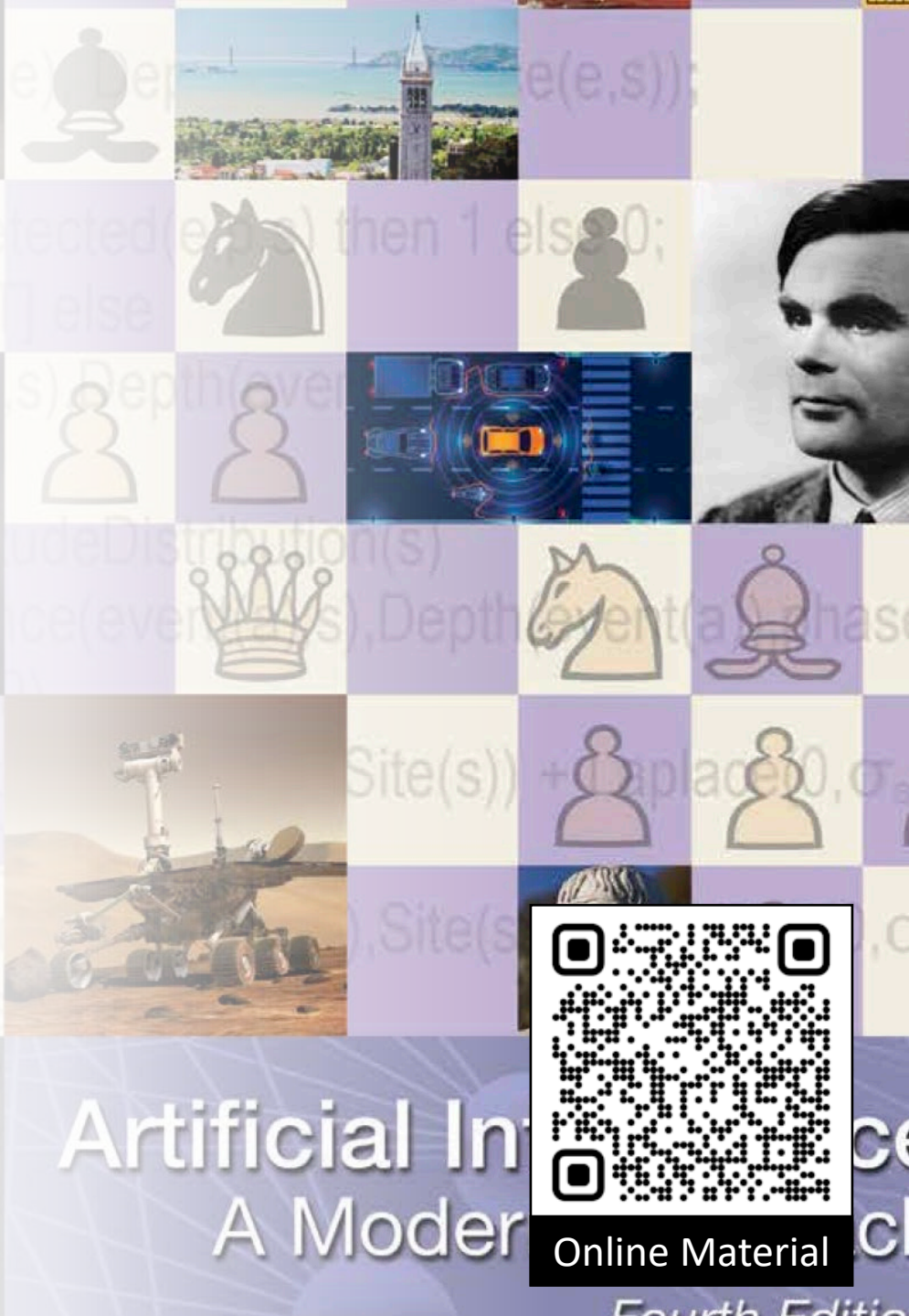
CS 5/7320
Artificial Intelligence

Introduction

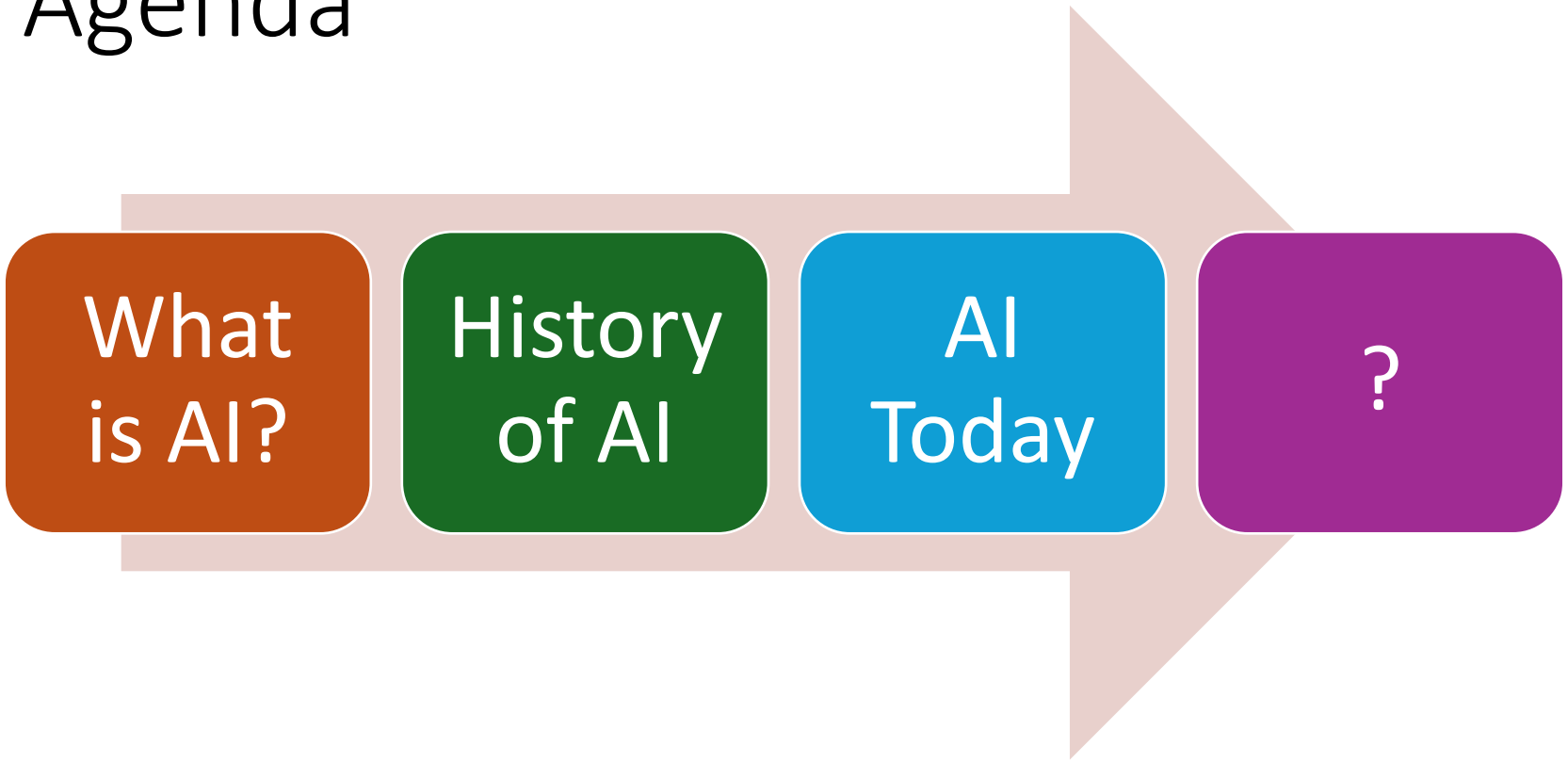
AIMA Chapters 1 + 27

Slides by Michael Hahsler

based on slides by Svetlana Lazepnik with figures and cover art from the AIMA textbook.



Agenda



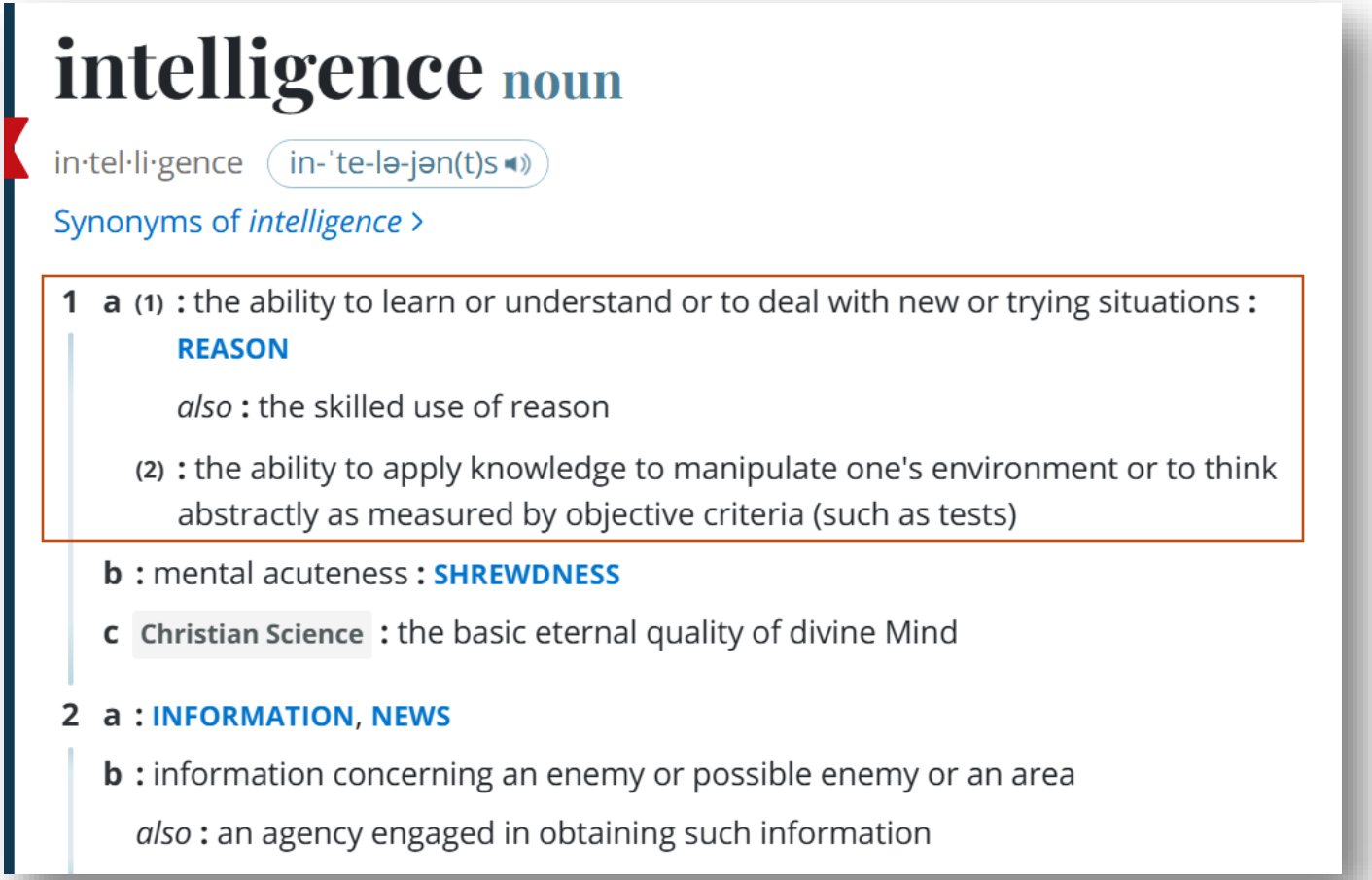
AI Ethics & Safety

What is AI?

ASIMO (**Advanced Step in Innovative Mobility**) is a humanoid robot created by Honda in 2000

What is Intelligence?

Definition from Merriam-Webster



intelligence noun

in·tel·li·gence in-ˈte-lə-jən(t)s

Synonyms of *intelligence* ›

1 **a** (1) **:** the ability to learn or understand or to deal with new or trying situations **:** **REASON**

also **:** the skilled use of reason

(2) **:** the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (such as tests)

b : mental acuteness **: SHREWDNESS**

c Christian Science **:** the basic eternal quality of divine Mind

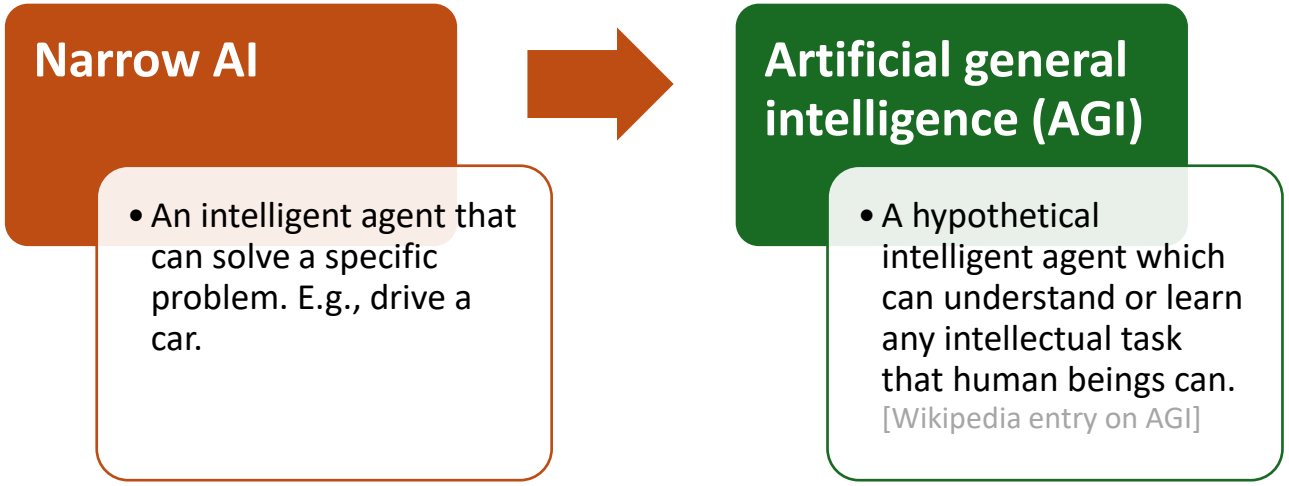
2 **a : INFORMATION, NEWS**

b : information concerning an enemy or possible enemy or an area

also **:** an agency engaged in obtaining such information

The Goal of AI

"Have machines solve problems that are challenging for humans."
We call such a machine an **intelligent agent.**



How can we achieve this? Create an agent that can:

Think like a human? | Act like a human? | Think rationally? | Act rationally?

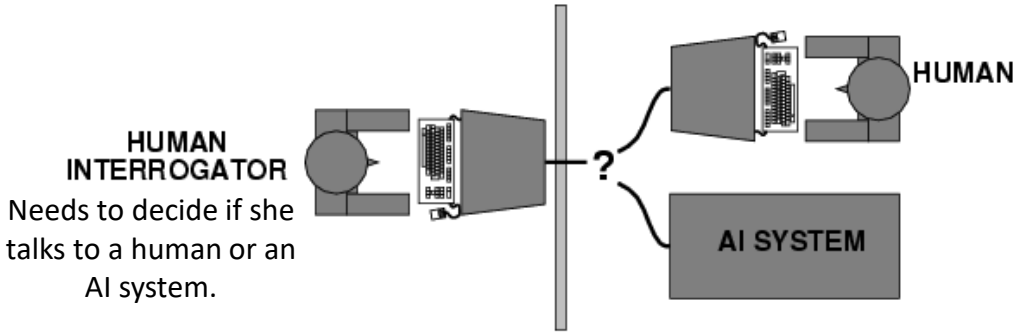
Think like a human? | Act like a human? | Think rationally? | Act rationally?

The brain as an information processing machine.	How to understand cognition as a computational process?	AI consciousness
• Requires scientific theories of how the brain works. **Note**: The brain does not work like artificial neural networks from ML!	• Introspection: try to think about how we think. • Predict the behavior of human subjects. • Image the brain, examine neurological data	• What does it mean that a machine is conscient/sentient? • How can we tell? (What do we do?)

Cognitive Sciences

Think like a human? | **Act like a human?** | Think rationally? | Act rationally?

- Alan Turing rejects the question “Can machines think?”
- The Turing Test tries to define what acting like a human means.



Alan Turing (1950) “Computing machinery and intelligence”

- What capabilities would a computer need to have to pass the Turing Test? These are still the core AI areas.
  - Natural language processing
  - Knowledge representation
  - Automated reasoning
  - Machine learning
- Turing predicted that by the year 2000, machines would be able to fool 30% of human judges for five minutes.
  ChatGPT (2023) is probably doing a least that!

Turing Test: Criticism

What are some potential problems with the Turing Test?

- Some human behavior is not intelligent.
- Some intelligent behavior may not be human.
- Human observers may be easy to fool.
  - A lot depends on expectations.
  - Anthropomorphic fallacy: humans tend to humanize things.
- Imitate intelligence without intelligence. E.g., the early chatbots ELIZA (1964) simulates a conversation using pattern matching.

Is passing the Turing test a good scientific goal?

- Engineering perspective: Imitating a human is not a good way to solve practical problems.
- We can create useful intelligent agents without trying to imitate humans.

Chinese Room Argument

Thought experiment by John Searle (1980): Imitate intelligence using rules.

Think like a human? | Act like a human? | **Think rationally?** | Act rationally?

- **Thinking Rationality**: Draw sensible conclusions from facts, logic and data.
- **Logic**: A chain of argument that always yield correct conclusions.
  E.g., "Socrates is a man; all men are mortal; therefore, Socrates is mortal."
- **Logic-based approach to AI**: Describe a problem in formal logic notation and apply general deduction procedures to solve it.
  Issues:
  - Describing real-world problems and knowledge using logic notation is hard.
  - Computational complexity of finding the solution.
  - Much intelligent or "rational" behavior in an uncertain world cannot be defined by simple logic rules.

Example: What about the logical implication

$$study\ hard \Rightarrow A\ in\ my\ AI\ course$$

Should it rather be

$$study\ hard\ AND\ be\ lucky\ AND\ \ldots \Rightarrow A\ in\ my\ AI\ course$$

Think like a human?

Act like a human?

Think rationally?

Act rationally?

Acting rational means to try to achieve the “best” outcome.

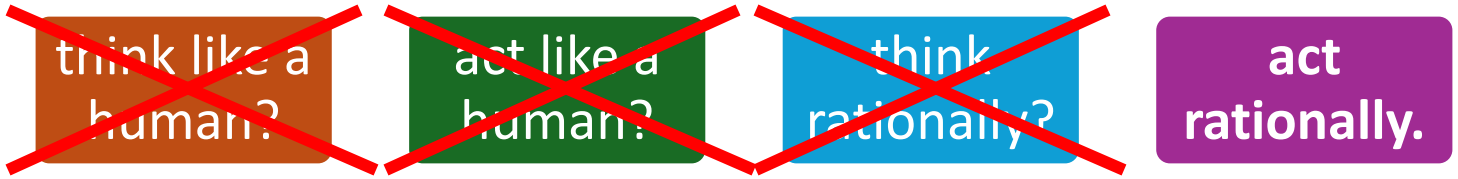
- Best means that we need to do **optimization**.
- The desirability of outcomes can be measured by the economic concept of **utility.**
- If there is uncertainty about achieving outcomes, then we need to maximizing the **expected utility.**

- Optimization has several advantages:
  - **Generality**: optimization is not limited to logical rules.
  - **Practicality**: can be adapted to many real-world problems.
  - **Well established**: existing solvers and methods for simulation and experimentation.
  - Avoids philosophy and psychology in favor of a **clearly defined objective**.

- **Bounded rationality:** In practice, expected utility optimization is subject to the agent’s knowledge and computational constraints. The agent needs to do the best with its knowledge and resources.

What type of AI do we cover in this course?

Create a **narrow AI agent** that can



That is, create a machines that acts in a way to solve a specific hard problem that traditionally would have been thought to require human intelligence.

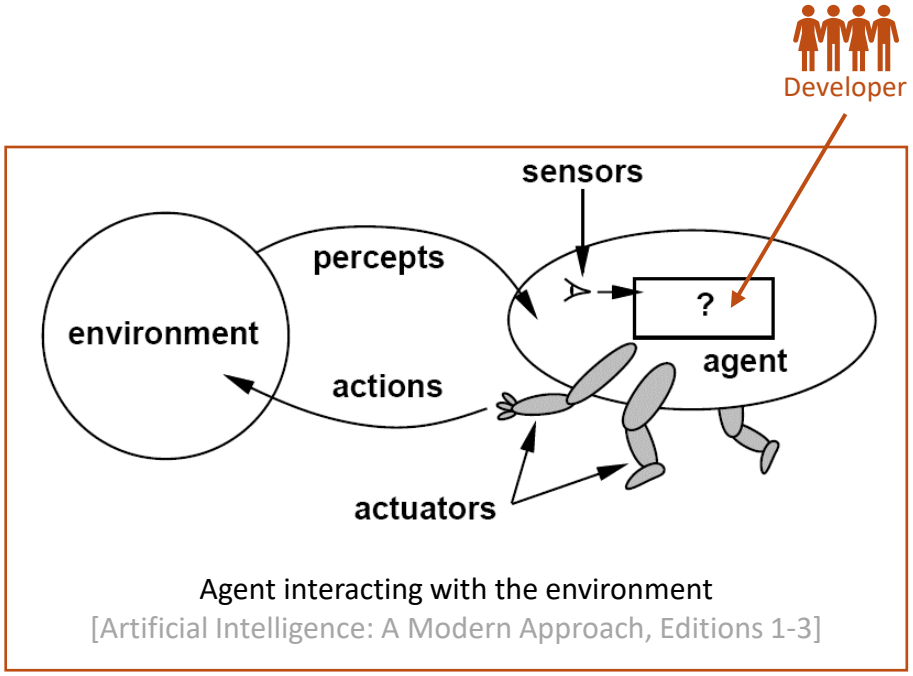
What are the Components of an Intelligent Agent?

Intelligent agents need to

- **Communicate** with the environment.
- **Represent knowledge**, **reason** and **plan** to achieve a desired **outcome**.

Optional

- **Learn** to improve performance**.**



Agent interacting with the environment
[Artificial Intelligence: A Modern Approach, Editions 1-3]

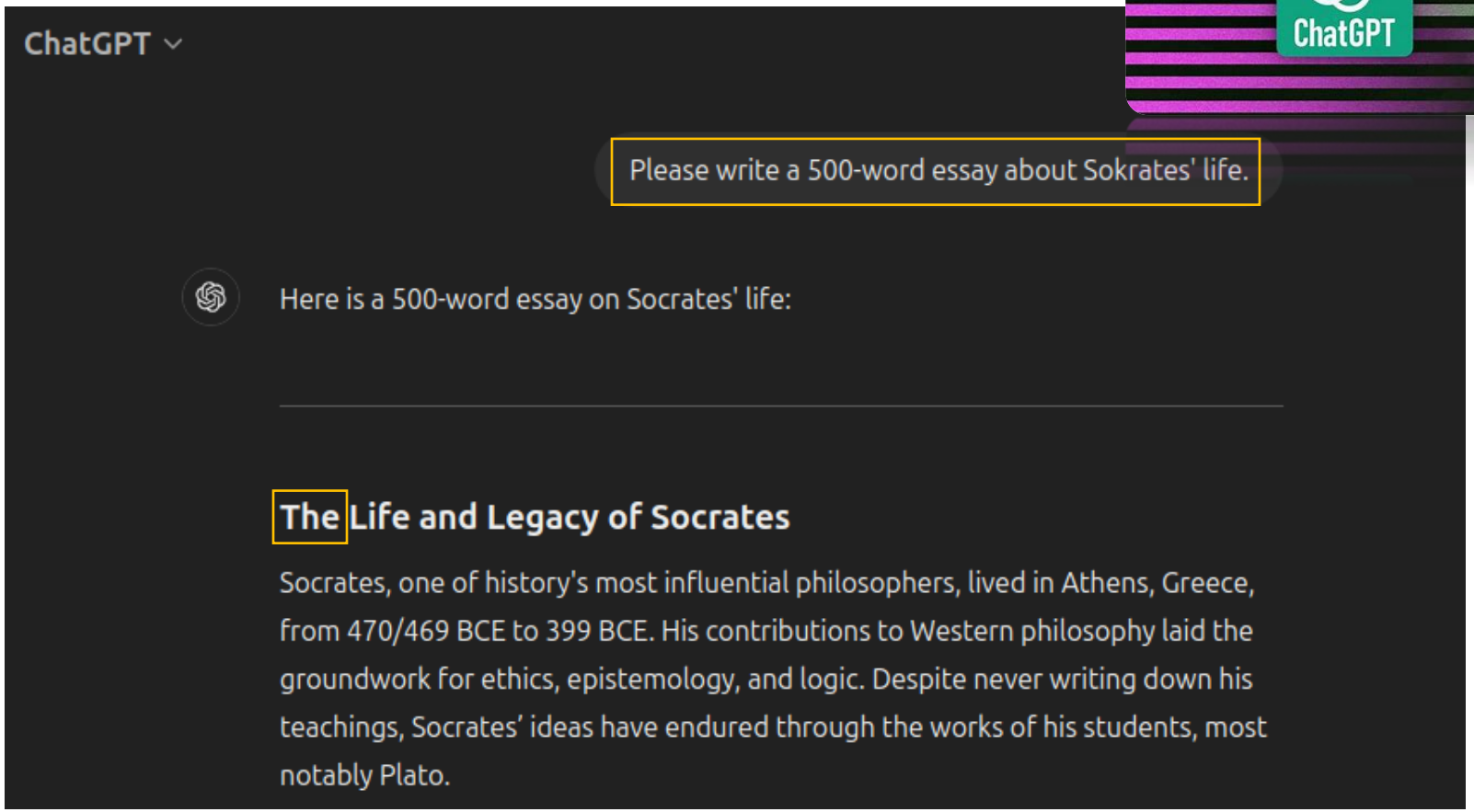
Example: Self-Driving Car

Percept: People crossing the street

Action: Stop the car

Objective: Reach the destination safely

Example: Homework and LLMs

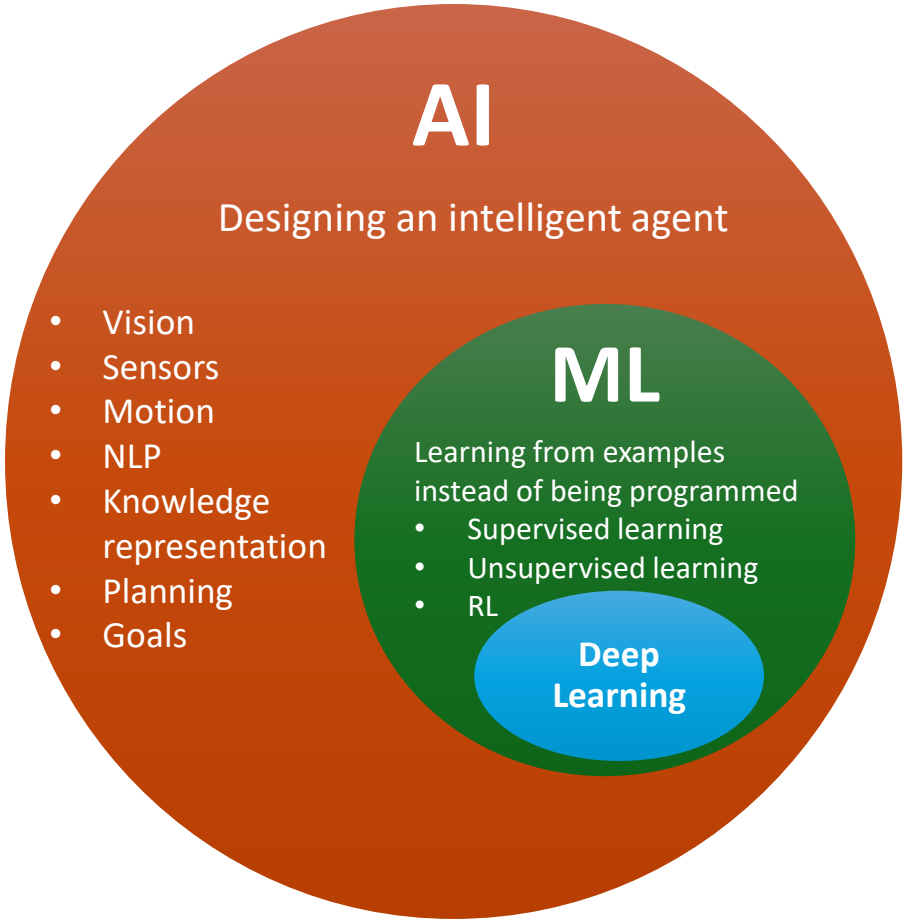


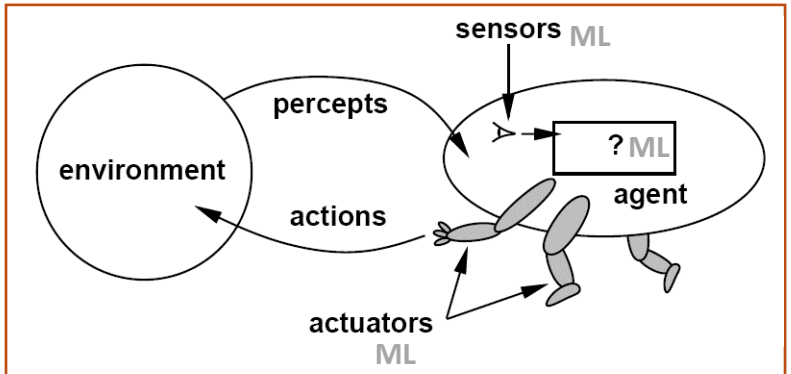
Percept: Your prompt

Action: Next most likely word… More words are created word-by-word.

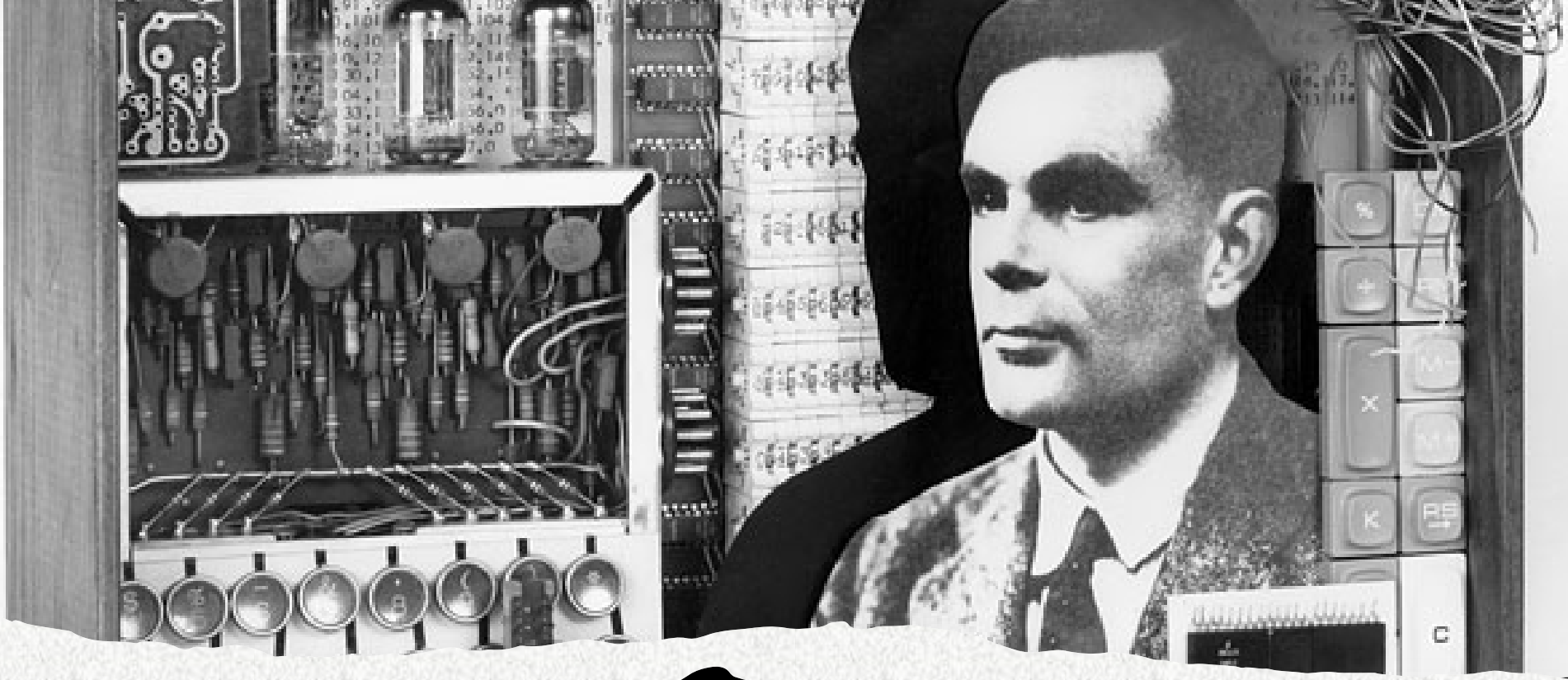
Objective: You may like a useful answer, but what is ChatGPT's objective?

Artificial Intelligence vs. Machine Learning

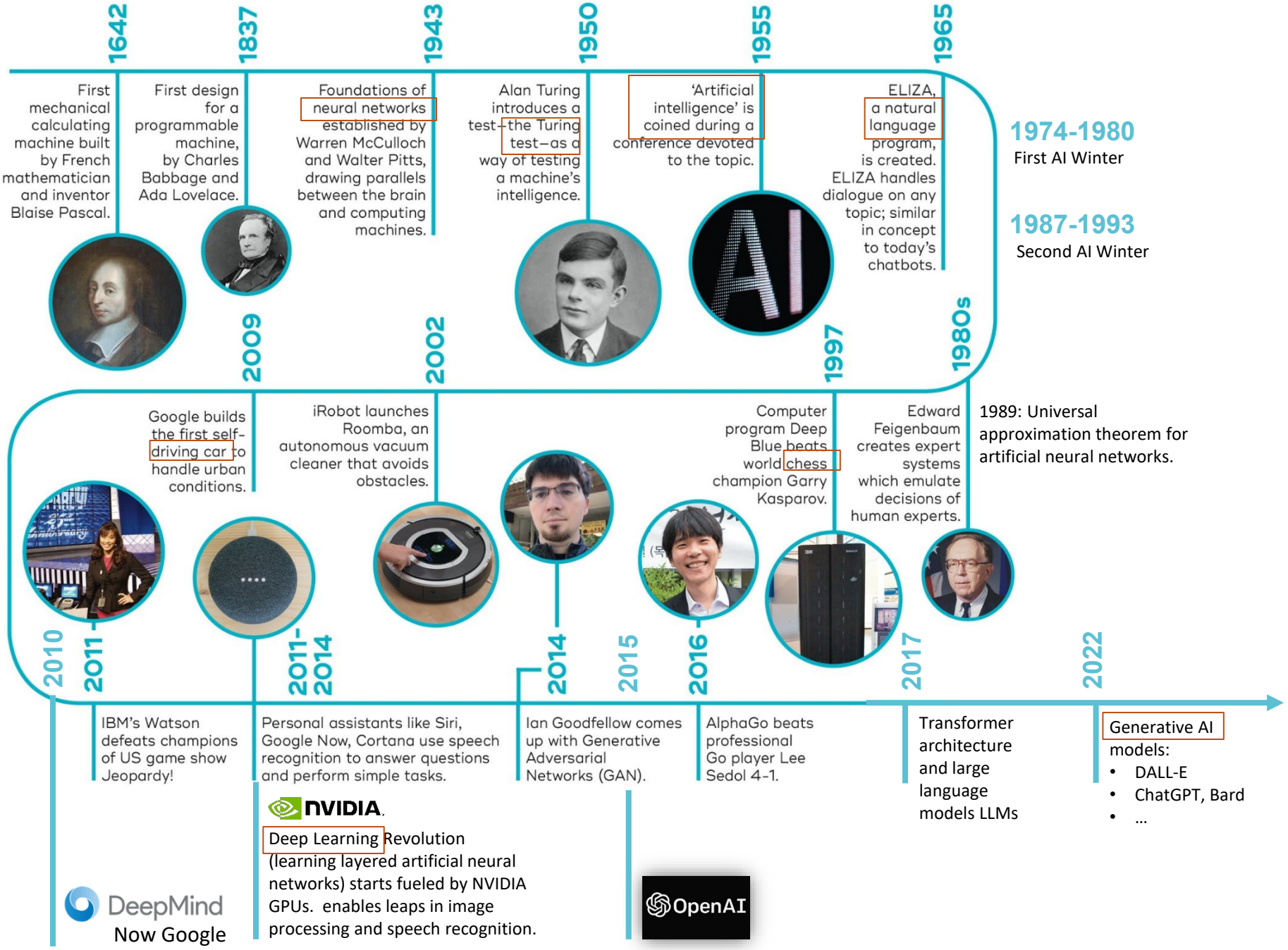




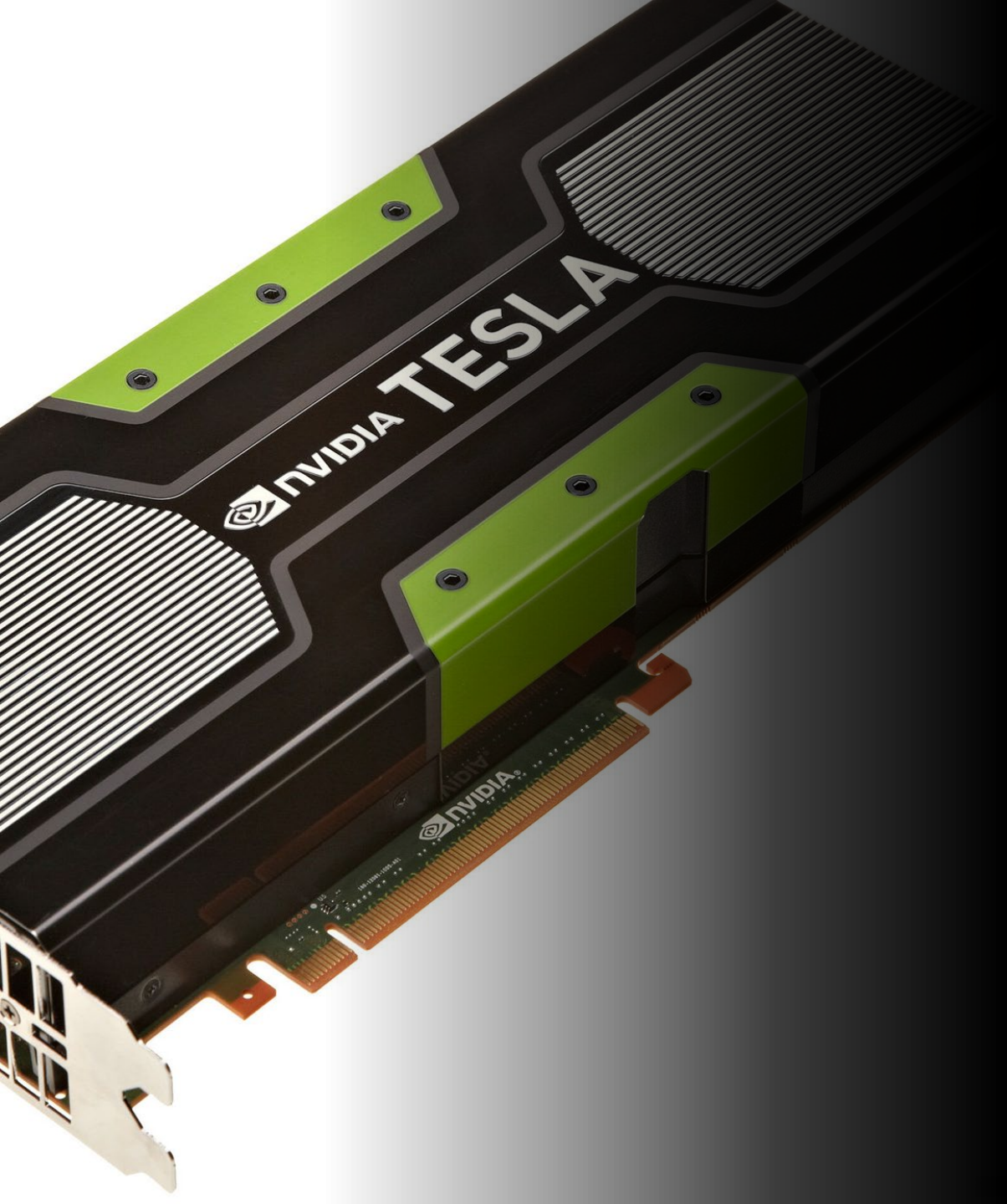
Agent interacting with the environment
[Artificial Intelligence: A Modern Approach, Editions 1-3]



The History of AI



Source: https://qbi.uq.edu.au/brain/intelligent-machines/history-artificial-intelligence + additions



What accounts for recent successes in AI?

- Faster computers and specialized hardware (GPUs).
- Lots of data (the Internet, text, sensors) and storage (cloud)
- Dominance of machine learning.
- New optimization methods (deep learning).

“Moravec’s Paradox”

Hans Moravec (1988): *“It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and* ***difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility****.”*

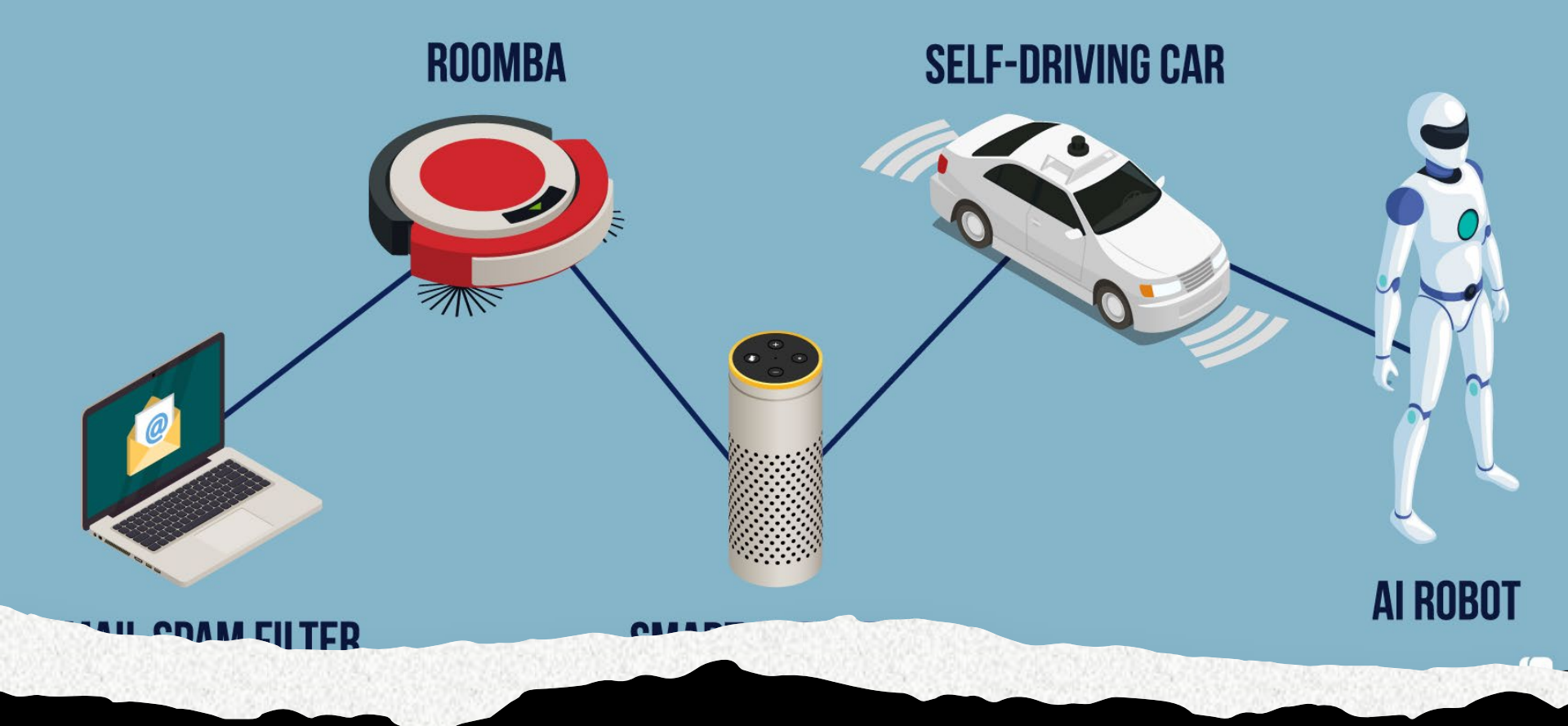
A teenager can learn how to drive in a few hours with very little input, but we still have no truly self-driving car.

https://www.newsweek.com/googles-new-two-legged-robot-future-warfare-429831

The AI Effect: AI gets no respect?

As soon as a machine gets good at performing some task, the task is no longer considered to require much intelligence.

- Calculating ability used to be prized – not anymore.
- Chess was thought to require high intelligence – now computers play at a super-human level.
- Learning once thought uniquely human - now machine learning is a well-developed discipline.
- Art? “Even a monkey can do this!”

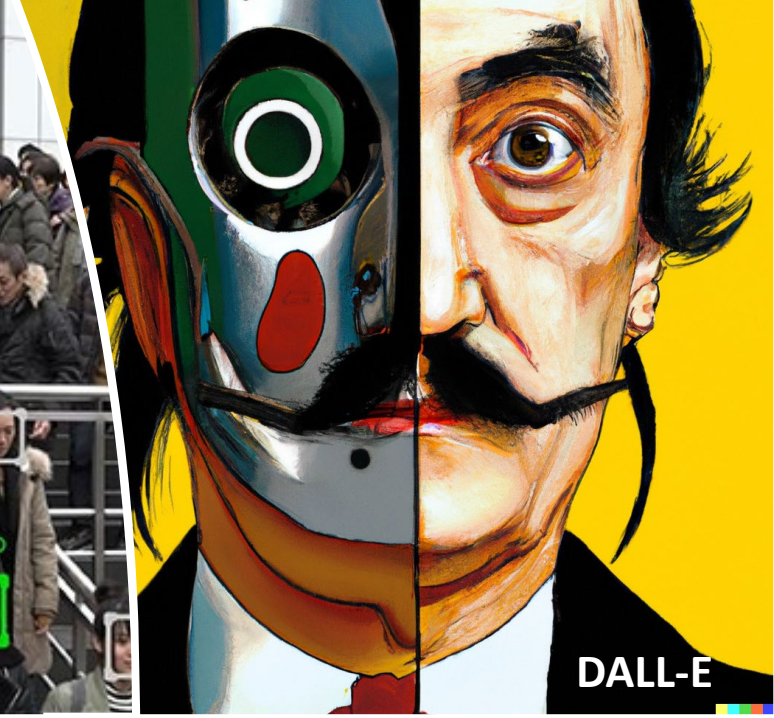


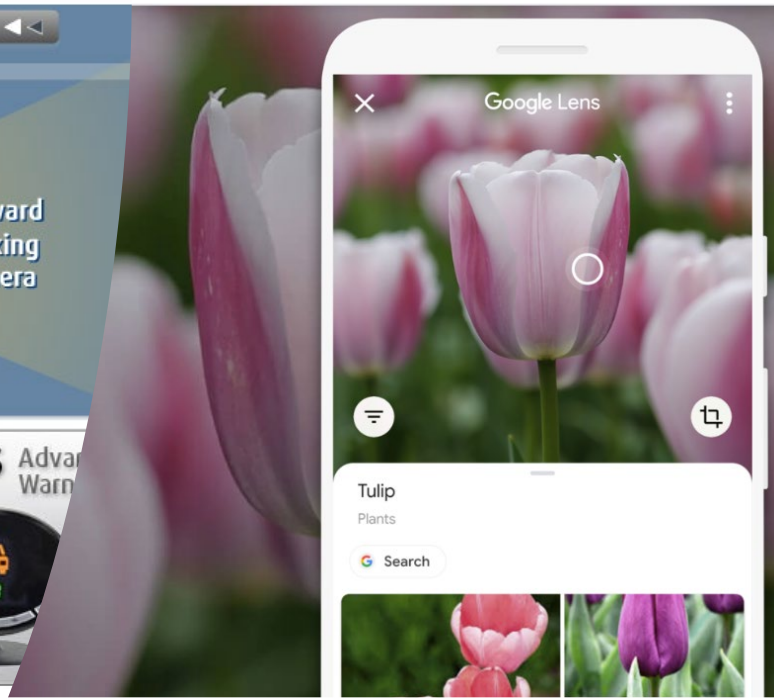
AI Today

Vision and Image Processing

- **OCR**: read license plates, handwriting recognition (e.g., mail sorting).
- **Face detection**: now standard for smart phone cameras.
- **Vehicle safety systems**
- **Visual search**
- **Image generation**

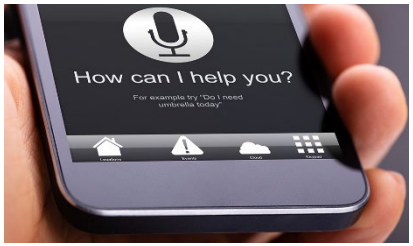
All these technologies operate now at superhuman performance.

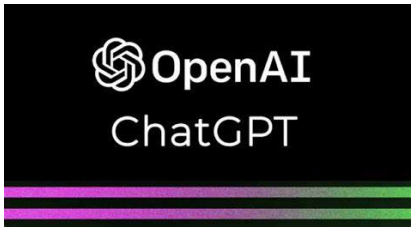


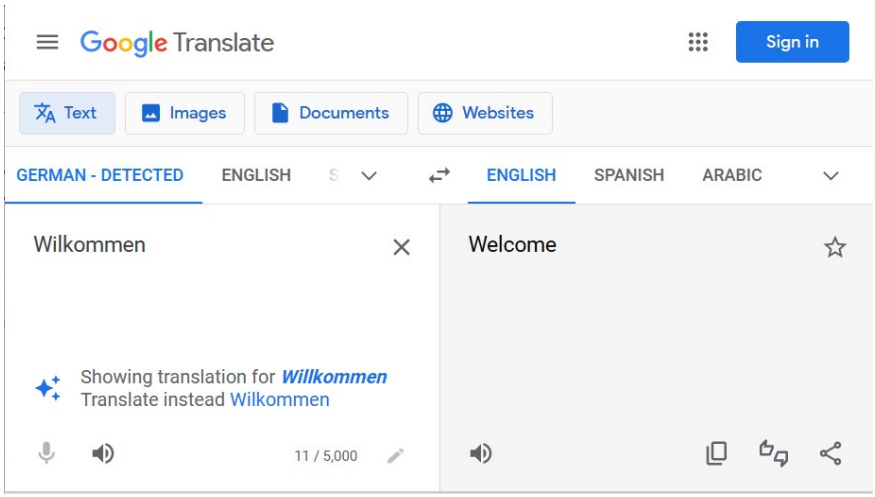


Natural Language Processing

- Text-to-speech
- Speech-to-text to detect voice commands
- Machine translation
- Text generation (Q/A systems) using Large Language Models







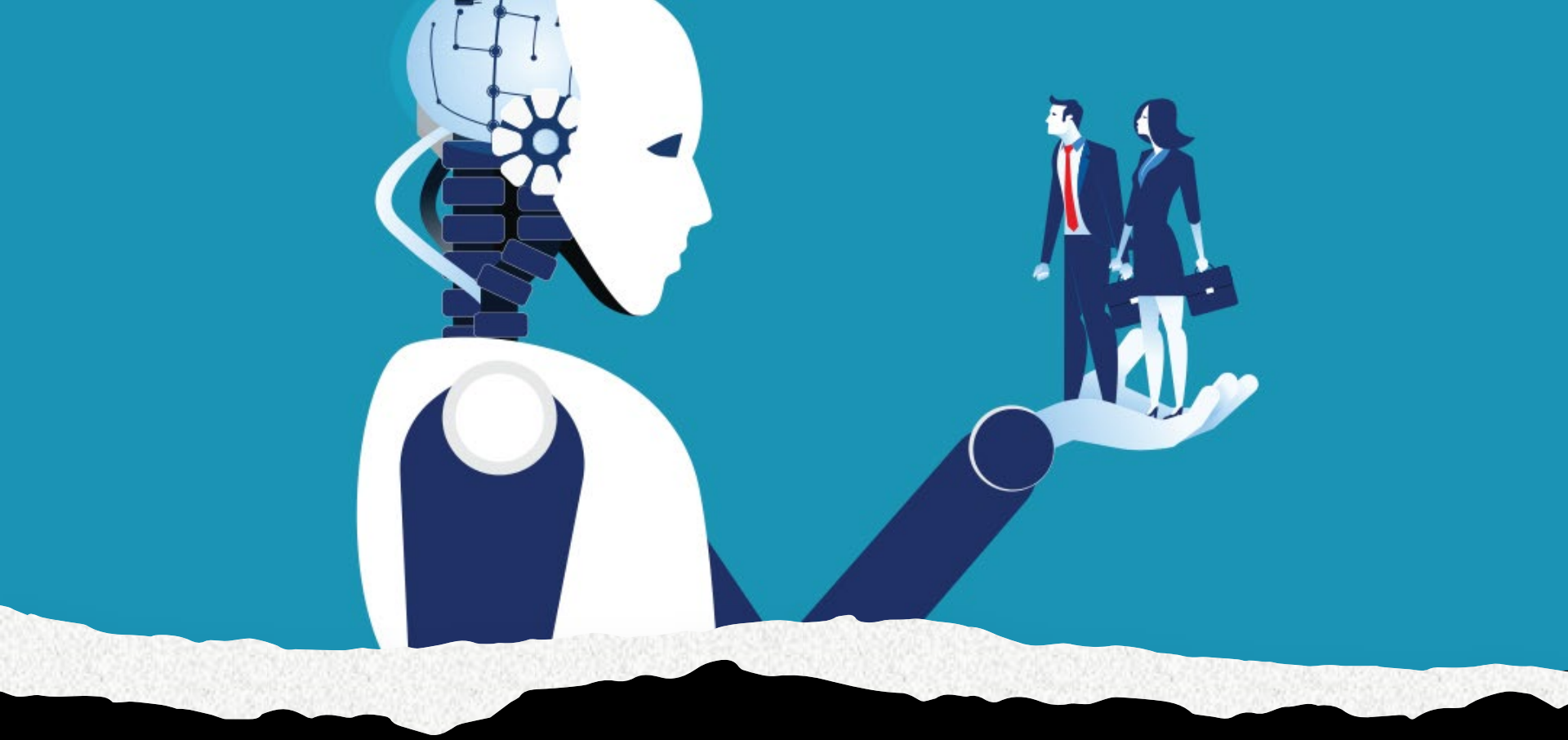
These technologies operate now with close to or even superhuman performance.

Humans use language to reason. Does that mean AI that can create high-quality text can reason?

Language understanding is still elusive!

Robotics

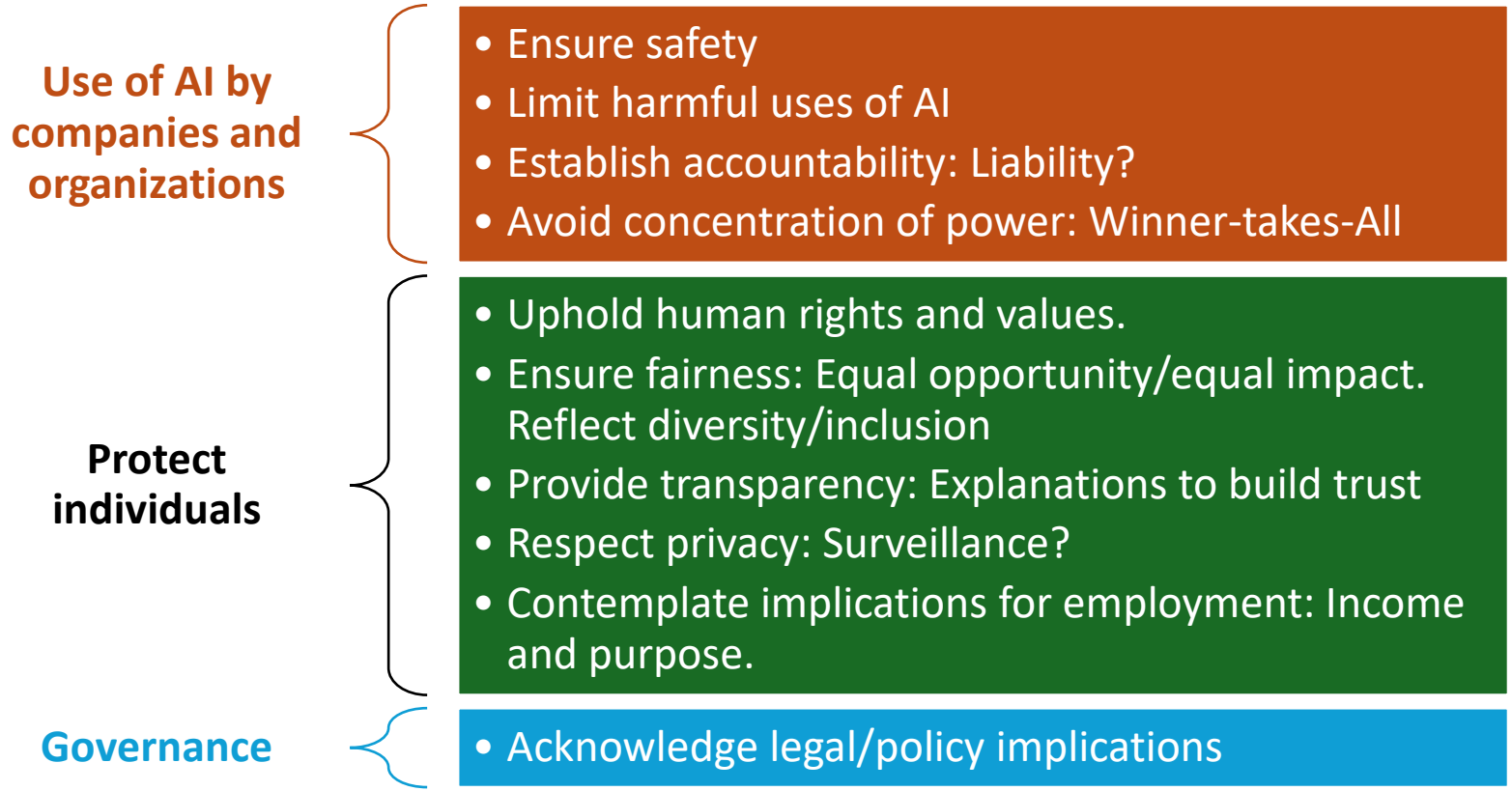
- Mars rovers
- Robot soccer
  - RoboCup
- Autonomous vehicles
  - DARPA Grand Challenge
  - Self-driving cars
- Drones
- Personal robotics
  - Humanoid robots
  - Robotic pets
  - Personal assistants?



AI Ethics & Safety

A new Frontier for Fairness and Freedom
AIMA Chapter 27

Commonly-Cited Safety and Ethics Principles



Next, we look at the implementation of these principles in different countries.

European Union

Has regulations since 2016 included in the General Data Protection Regulation (GDPR)

Art. 22 GDPR – Automated individual decision-making, including

California Consumer Privacy Act

California's CCPA was not modeled after the GDPR

Art. 22 GDPR

2016

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   (c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

In the United States

Bill Introduced was 2019
Did not receive a vote in Congress.

116TH CONGRESS
1ST SESSION

H. R. 2231

To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.

(2) AUTOMATED DECISION SYSTEM IMPACT ASSESSMENT.—The term "automated decision system impact assessment" means a study evaluating an automated decision system and the automated decision system's development process, including the design and training data of the automated decision system, for impacts on accuracy, fairness, bias, discrimination, privacy, and security that includes, at a minimum—

(A) a detailed description of the automated decision system, its design, its training, data, and its purpose;

(B) an assessment of the relative benefits and costs of the automated decision system in light of its purpose, taking into account relevant factors, including—

(i) data minimization practices;

(ii) the duration for which personal information and the results of the automated decision system are stored;

(iii) what information about the automated decision system is available to consumers;

(iv) the extent to which consumers have access to the results of the automated decision system and may correct or object to its results; and

(v) the recipients of the results of the automated decision system;

(C) an assessment of the risks posed by the automated decision system to the privacy or security of personal information of consumers and the risks that the automated decision system may result in or contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers; and

(D) the measures the covered entity will employ to minimize the risks described in subparagraph (C), including technological and physical safeguards.

European Union Study

2019

A governance framework for algorithmic accountability and transparency

This study develops policy options for the governance of algorithmic transparency and accountability, based on an analysis of the social, technical and regulatory challenges posed by algorithmic systems. Based on a review and analysis of existing proposals for governance of algorithmic systems, a set of four policy options are proposed, each of which addresses a different aspect of algorithmic transparency and accountability: 1. awareness raising: education, watchdogs and whistleblowers; 2. accountability in public-sector use of algorithmic decision-making; 3. regulatory oversight and legal liability; and 4. global coordination for algorithmic governance.

https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262

2021

Background

Google has long championed AI. Our research teams are at the forefront of AI development, and we've seen firsthand how AI can enable massive increases in performance and functionality. AI has the potential to deliver great benefits for economies and society — from improving energy efficiency and more accurately detecting disease, to increasing the productivity of businesses of all sizes. Harnessed appropriately, AI can also support fairer, safer and more inclusive and informed decision-making. We are keen to ensure that everyone and every business can benefit from the opportunities that AI creates.

AI will have a significant impact on society for many years to come. That's why we established our AI Principles (including applications we will not pursue)[1] to guide Google teams on the responsible development and use of AI. These are backed by the operational processes and structures necessary to ensure they are not just words but concrete standards that actively impact our research, products and business decisions to ensure trustworthy and effective AI application.

But while self-regulation is vital, it is not enough. Balanced, fact-based guidance from governments, academia and civil society is also needed to establish boundaries, including in the form of regulation. As our CEO Sundar Pichai has noted, AI is too important not to regulate. The challenge is to do so in a way that is proportionately tailored to mitigate risks

Source: https://ai.google/static/documents/recommendations-for-regulating-ai.pdf, 2021

Administration Priorities The Record

US White House Executive Order 14110

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

2023

Some important points:

- Artificial Intelligence must be **safe and secure**.
- Promoting **responsible innovation, competition, and collaboration**
- Americans' **privacy, civil liberties and labor rights** must be protected.

January 2025: This Executive Order is now revoked.

Fairness: Algorithmic Bias

"**Algorithmic bias** describes systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others." [Wikipedia]

Pre-existing bias

Social and institutional norms influence design and training data choices.

Example: Evaluate job applicants for a job which is historically almost exclusively held by males.

Technical bias

Limitations of a program or computational power.

Example: instead of a random sample, the program uses the first n data points.

Emergent bias

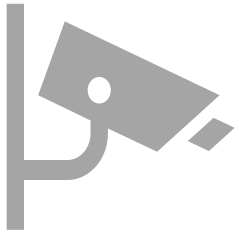
Use and reliance on algorithms across new or unanticipated contexts.

Example: Use a large language model to help judges decide on the length of prison sentence.

Types of AI Safety

"Prevent accidents, misuse, or other harmful consequences of AI."

AI Testing



Monitoring AI

Adversarial robustness

How should this be ensured?

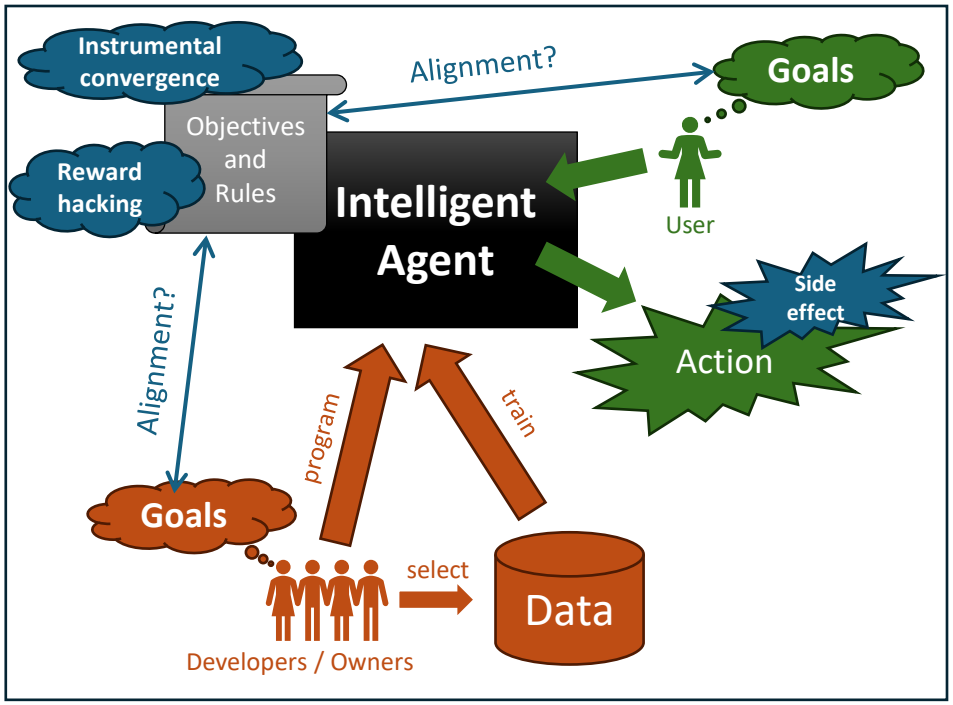
- Corporate self-regulation
- Private watchdogs
- Government action
- International treaties

Is it reasonable to assume that a superintelligent AI can get around being tested and monitored?

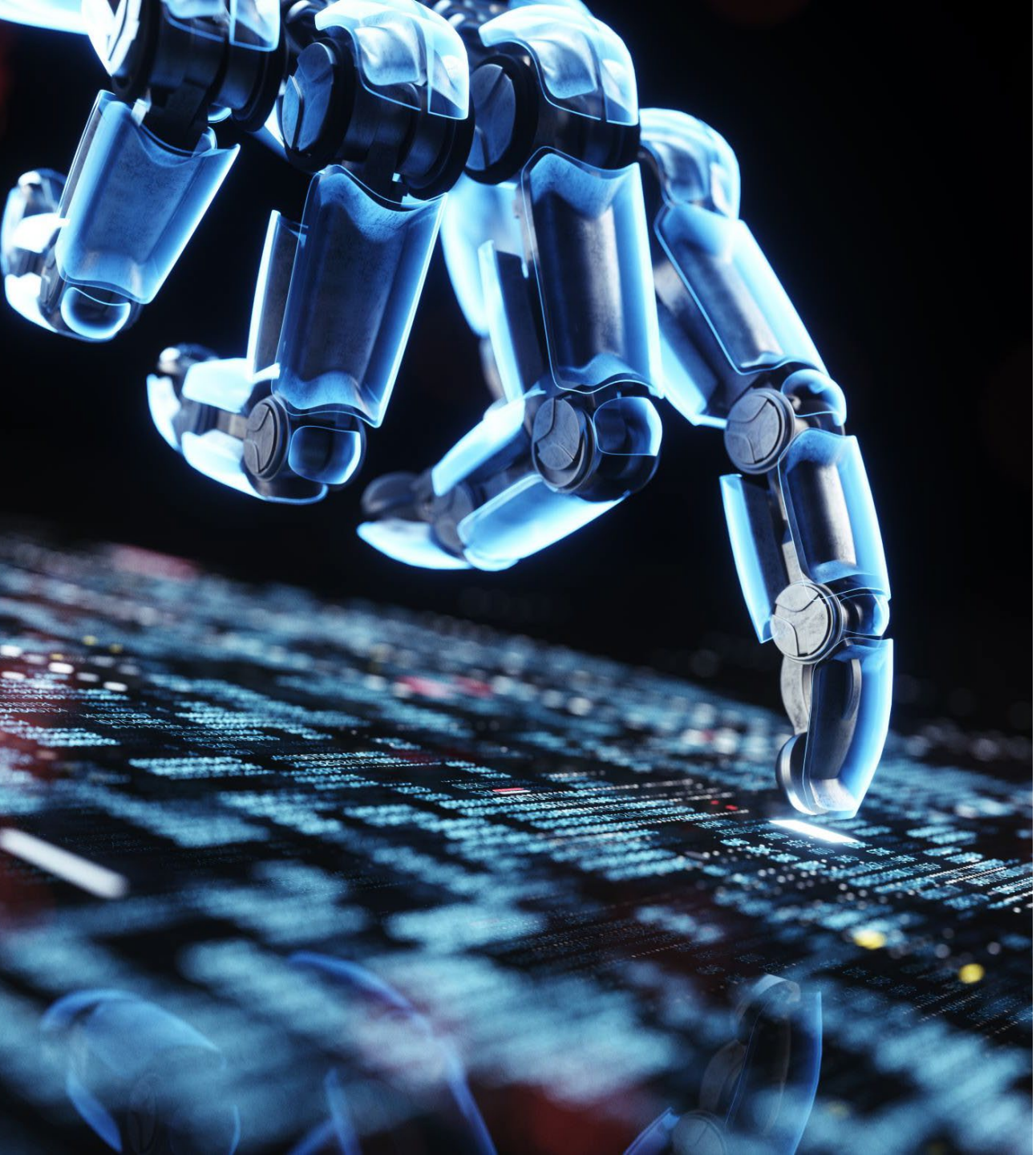
AI Safety and Optimizers

Intelligent Agents are “optimizers!”

- **Goal/reward alignment**: How do we specify a robust objective function? Whose objectives are used?
- **Reward hacking:** The AI learns to exploit unintended side effects to get a high “score” without solving the objective. AI needs to follow social norms.
- **Instrumental convergence**: All intelligent agents will pursue common subgoals like the need for more power to get better at reaching its objectives. How will this need be balanced with human’s needs?



Credit: Terminator 3: Rise of the Machines. Warner Bros.



Outlook

AI is a technology that is on the verge of significant leaps…

- New technologies always had a **profound impacted** on the way we live and work (e.g., electricity, the internet, mobile communication).
- We can expect unprecedented gains in productivity from better **narrow AI**.
- New technologies always also present **dangers** and need to be regulated.

This course will introduce simple techniques to create intelligent agents.