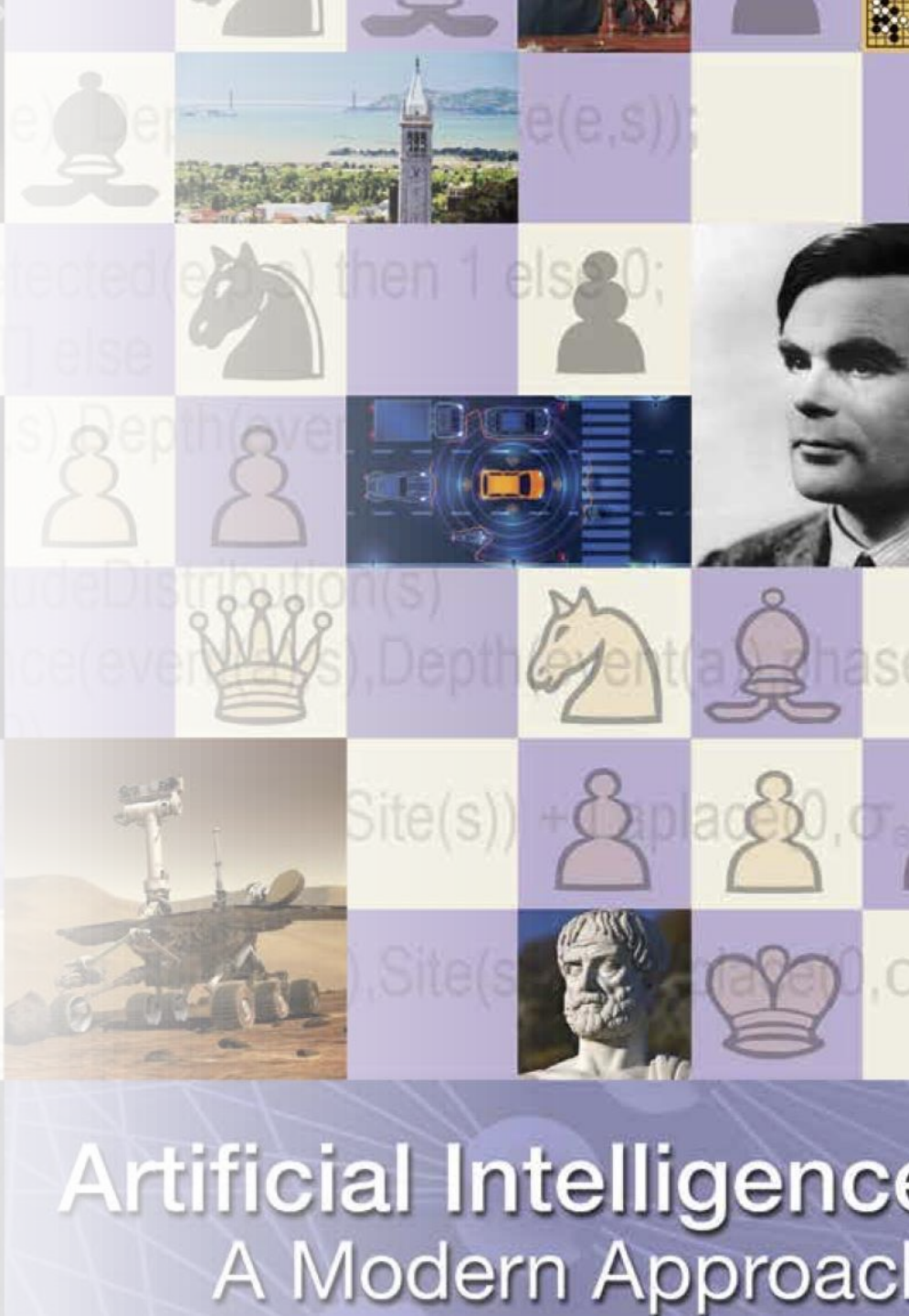
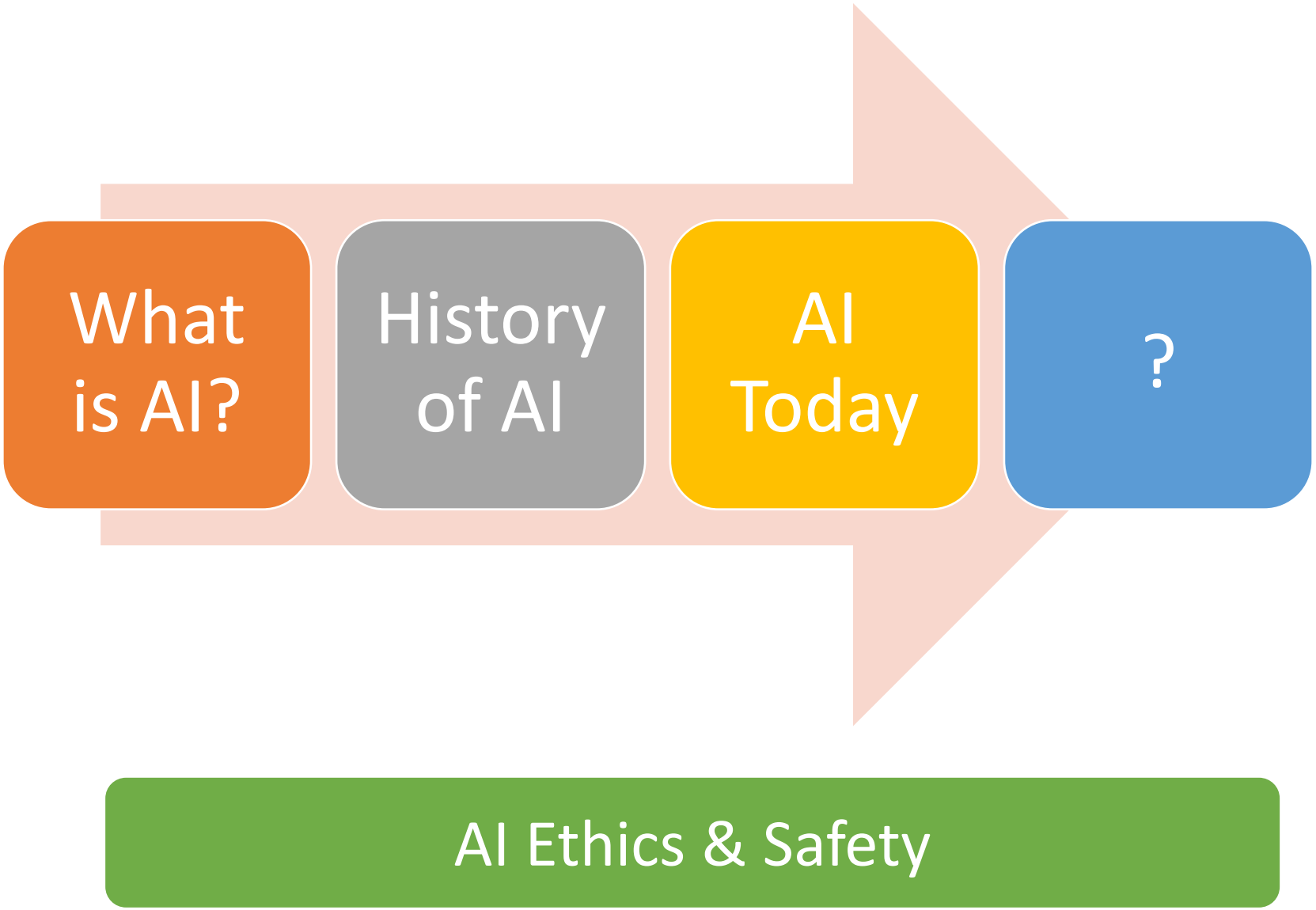
CS 5/7320 Artificial Intelligence

Introduction

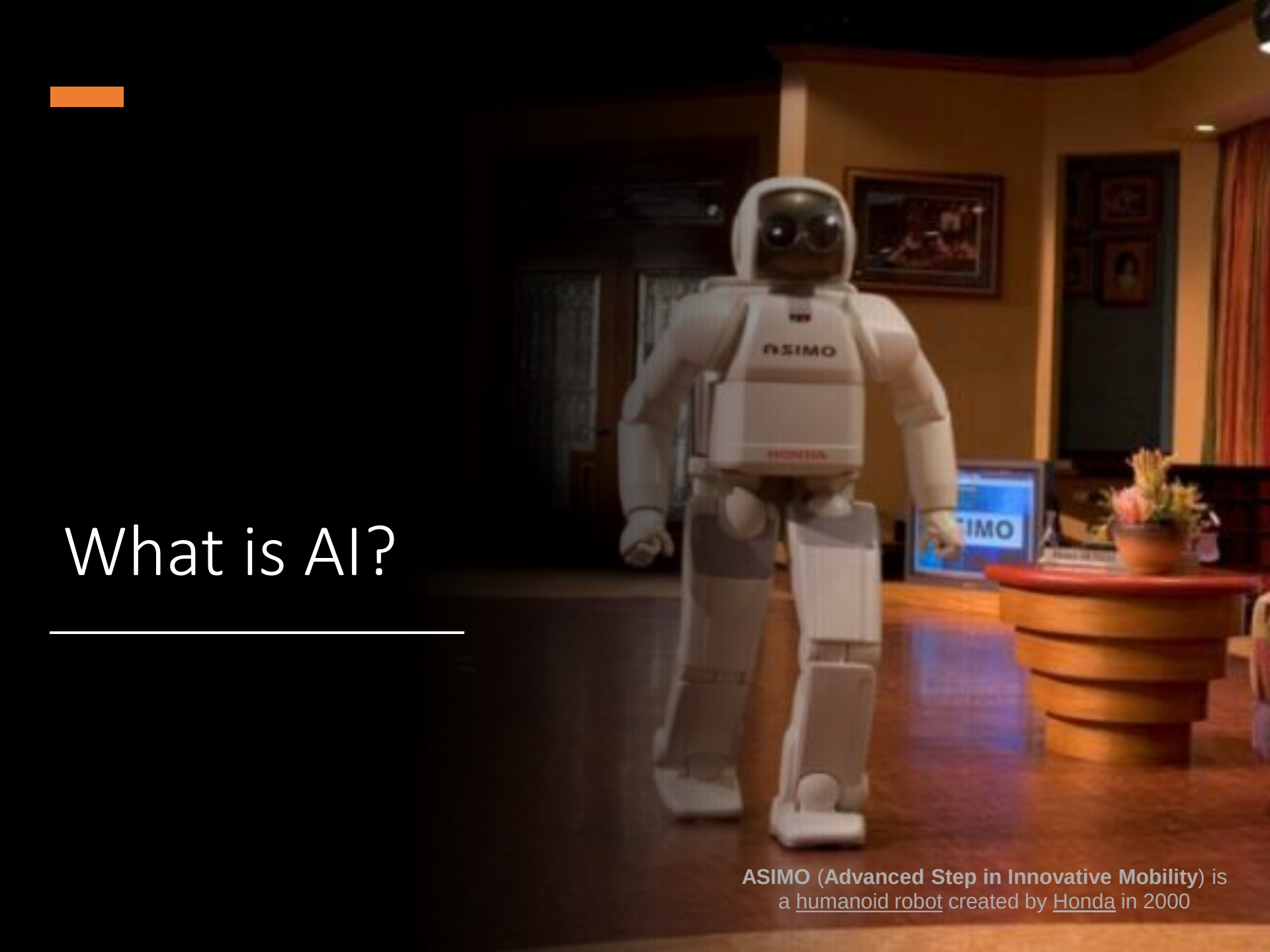
AIMA Chapter 1

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




What is AI?
History of AI
AI Today
?
AI Ethics & Safety

What is AI?



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

What is it the Goal of AI?

"Have machines solve problems that a challenging for humans."

- **Narrow AI** focuses on intelligent agents to solve a specific subproblem.
- An **artificial general intelligence (AGI)** is a hypothetical intelligent agent which can understand or learn any intellectual task that human beings or other animals can. [Wikipedia entry on AGI]

How do we achieve this?

Create an agent that

thinks like a human?	acts like a human?	thinks rationally?	acts rationally?

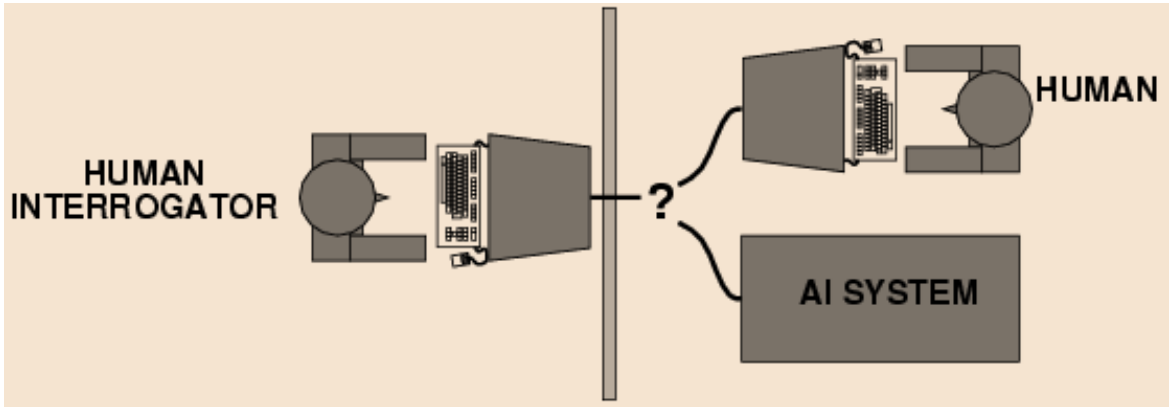
Thinking Like a Human

The brain as an information processing machine.	How to understand cognition as a computational process?
• 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

= Cognitive Sciences

Acting Like a Human

- Alan Turing (1950) "Computing machinery and intelligence"
- The Turing Test tries to define what acting like a human means



- What capabilities would a computer need to have to pass the Turing Test?
  - 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 in 2023 is probably doing a lot better than 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: 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.

What about modern chatbots like ChatGPT?

Thinking Rationally

- Idealized or “right” way of thinking.
- **Logic:** Patterns of argument that always yield correct conclusions when supplied with correct premises
  - “Socrates is a man; all men are mortal; therefore, Socrates is mortal.”
  - Beginning with Aristotle (385 BC), philosophers and mathematicians have attempted to formalize the rules of logical thought.
- ***Logic*-based approach to AI:** Describe problem in formal logical notation and apply general deduction procedures to solve it.
- Problems with the logic-based approach to AI
  - Describing real-world problems and knowledge in logical notation is hard.
  - Computational complexity of finding the solution.
  - A lot of intelligent or “rational” behavior in an uncertain world cannot be defined by simple rules.

What about the logical implication

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

Should it be

$$study\ hard\ AND\ be\ lucky \Rightarrow A\ in\ AI$$

Acting Rationally

A rational agent acts to achieve the best expected outcome:

- Goals are application-dependent and are expressed in terms of the **utility of outcomes.**
- Being rational means acting to **maximizing your expected utility.** Expectation means that different outcomes are possible (probabilities).
- In practice, utility optimization is subject to the agent's knowledge and computational constraints (**bounded rationality** or bounded optimality).

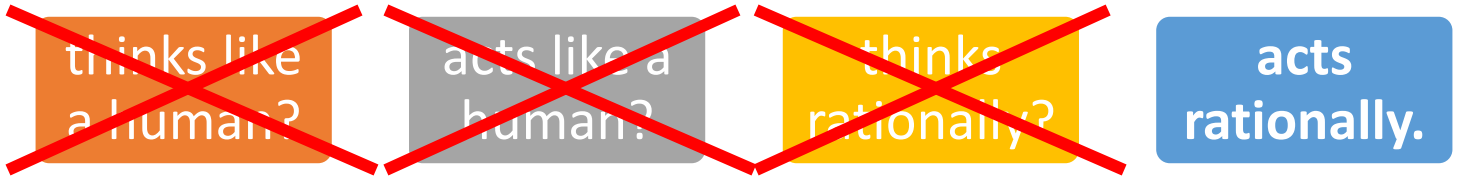
Acting Rationally

Advantages of the “expected utility maximization” formulation

- **Generality**: an optimization that goes beyond explicit reasoning with rules.
- **Practicality**: can be adapted to many real-world problems.
- Amenable to good scientific and engineering methodology including simulation and experimentation.
- Only concerns the decisions/actions that are made, not the cognitive process behind them. Avoids philosophy and psychology in favor of a clearly defined objective.

What type of AI do we cover in this course?

Create a narrow AI agent that



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

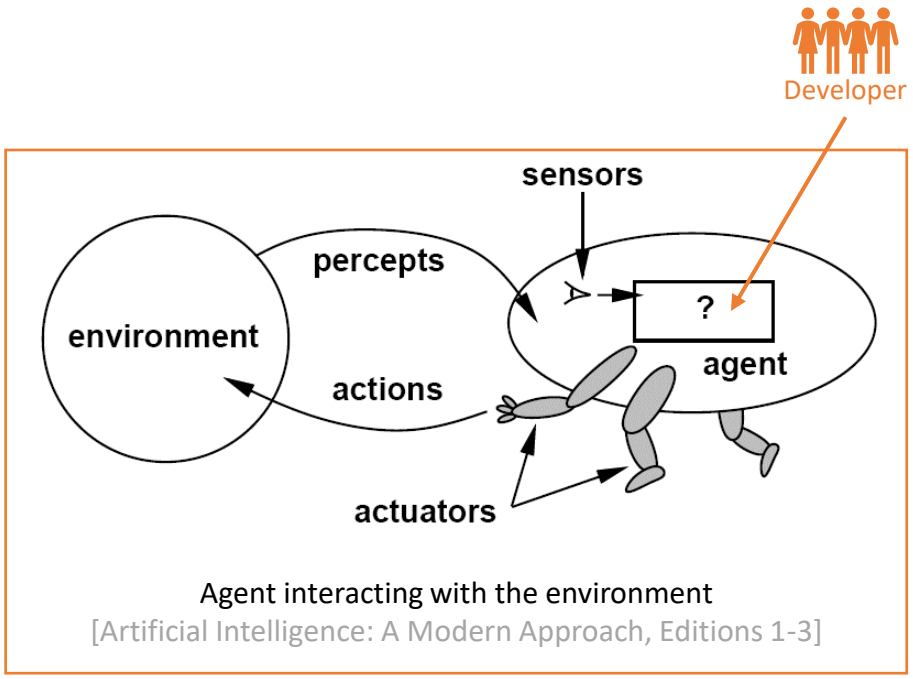
What are the Components of an Intelligent Agent?

Intelligent agents need to

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

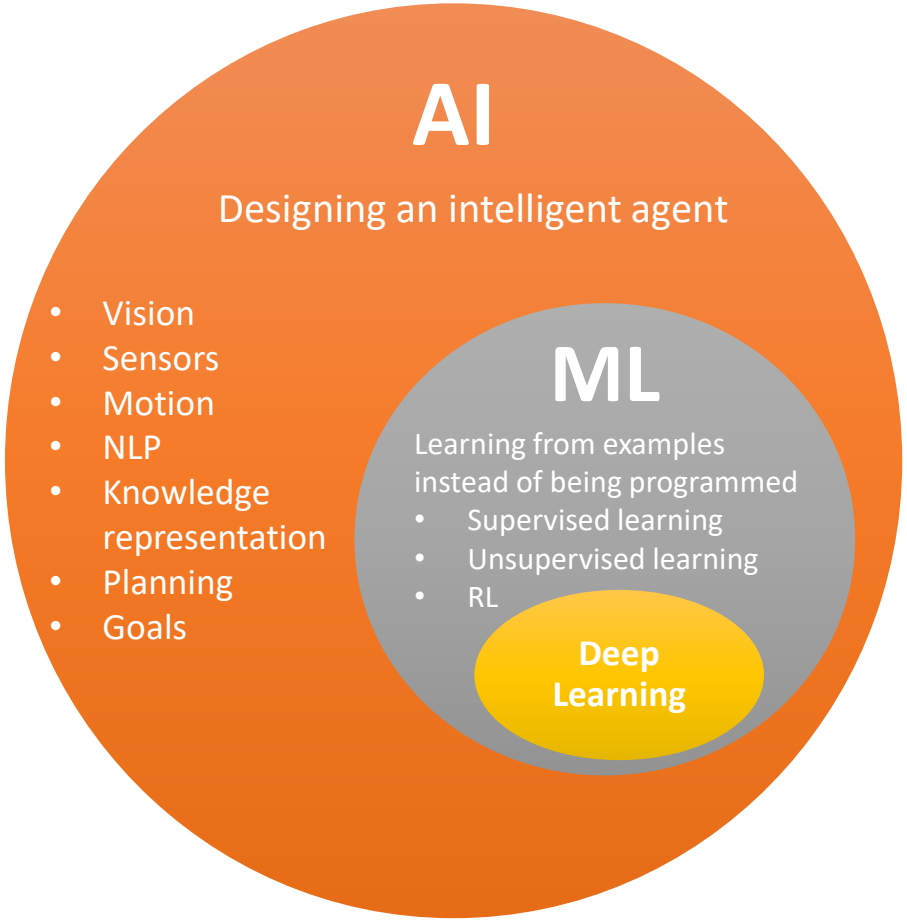
Optional

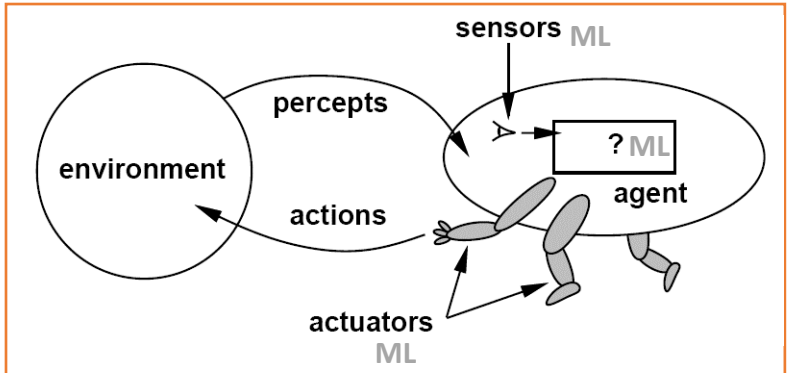
- Learn



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

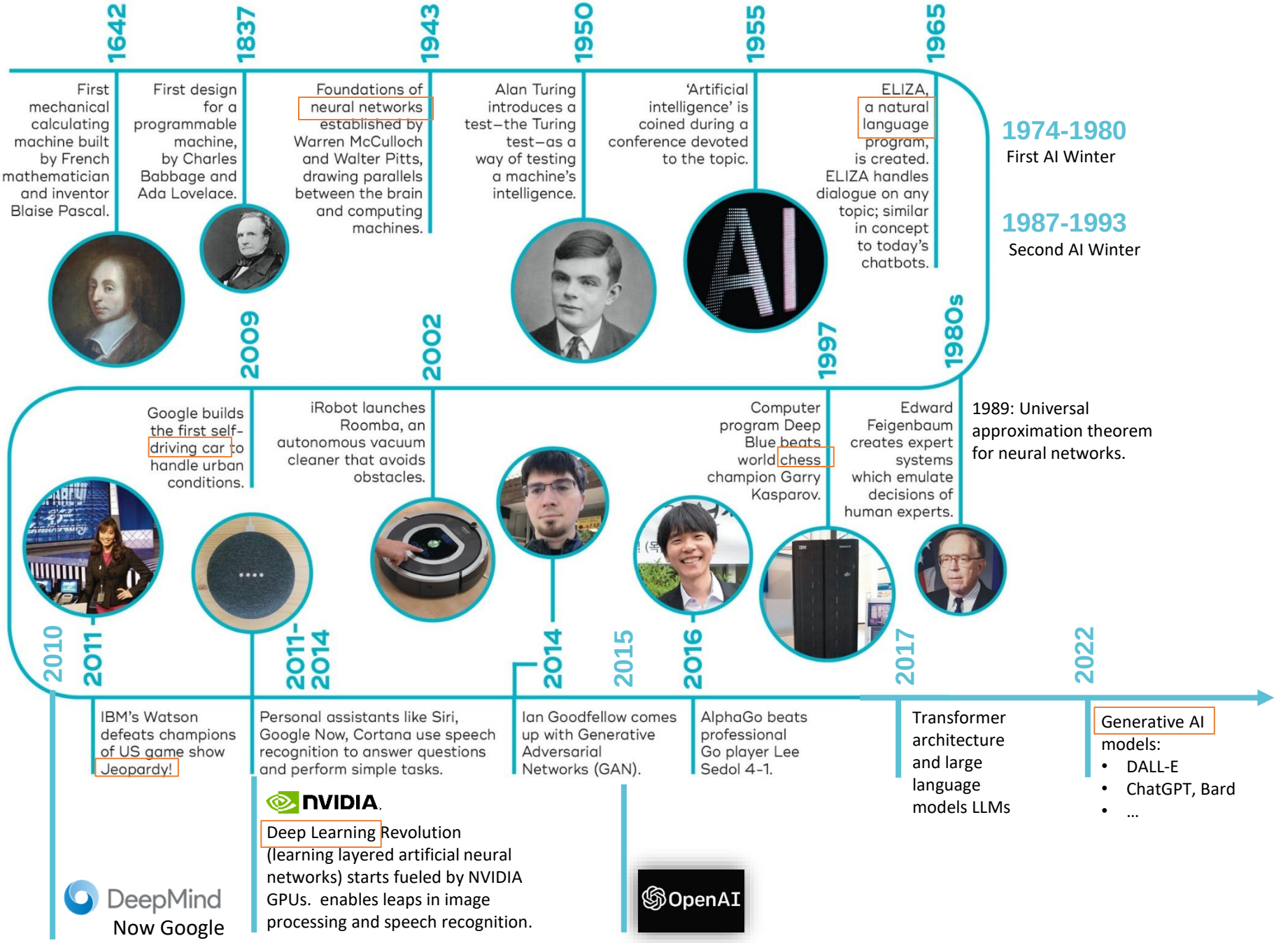
Machine Learning vs. Artificial Intelligence





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

AI is harder than originally thought

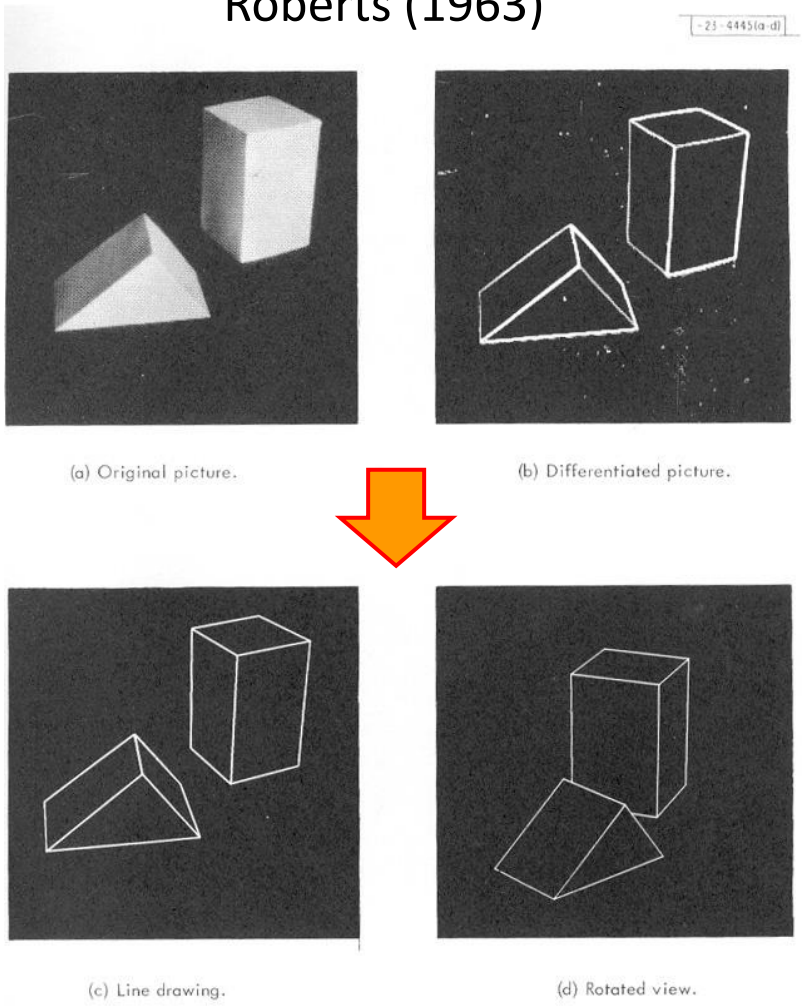
Herbert Simon, 1957

"It is not my aim to surprise or shock you--- but … there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until---in a visible future---the range of problems they can handle will be coextensive with the range to which human mind has been applied. ***More precisely: within 10 years a computer would be chess champion, and an important new mathematical theorem would be proved by a computer****."*

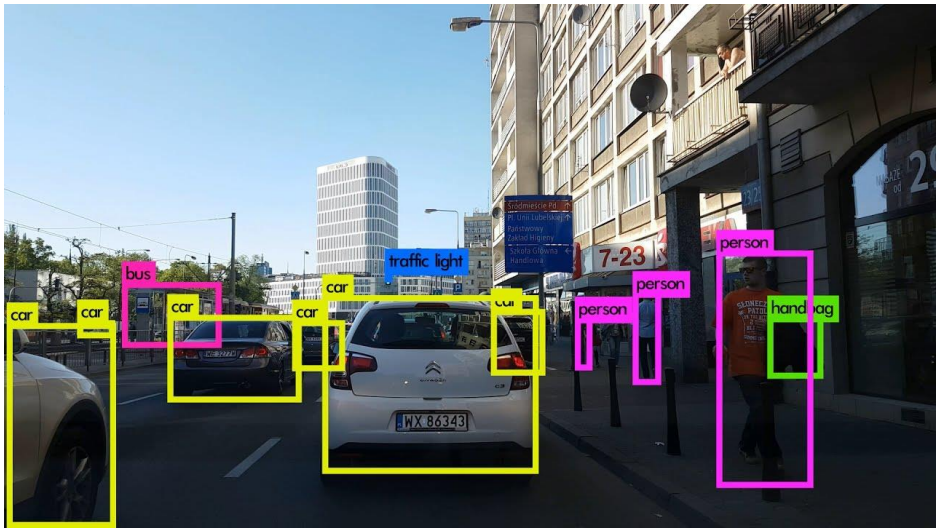
Simon's prediction came true --- but 40 years later instead of 10

From Blocks World to Modern Object Recognition

Roberts (1963)



Now



This is a lot harder!
But we can do it now….

“Moravec’s Paradox”

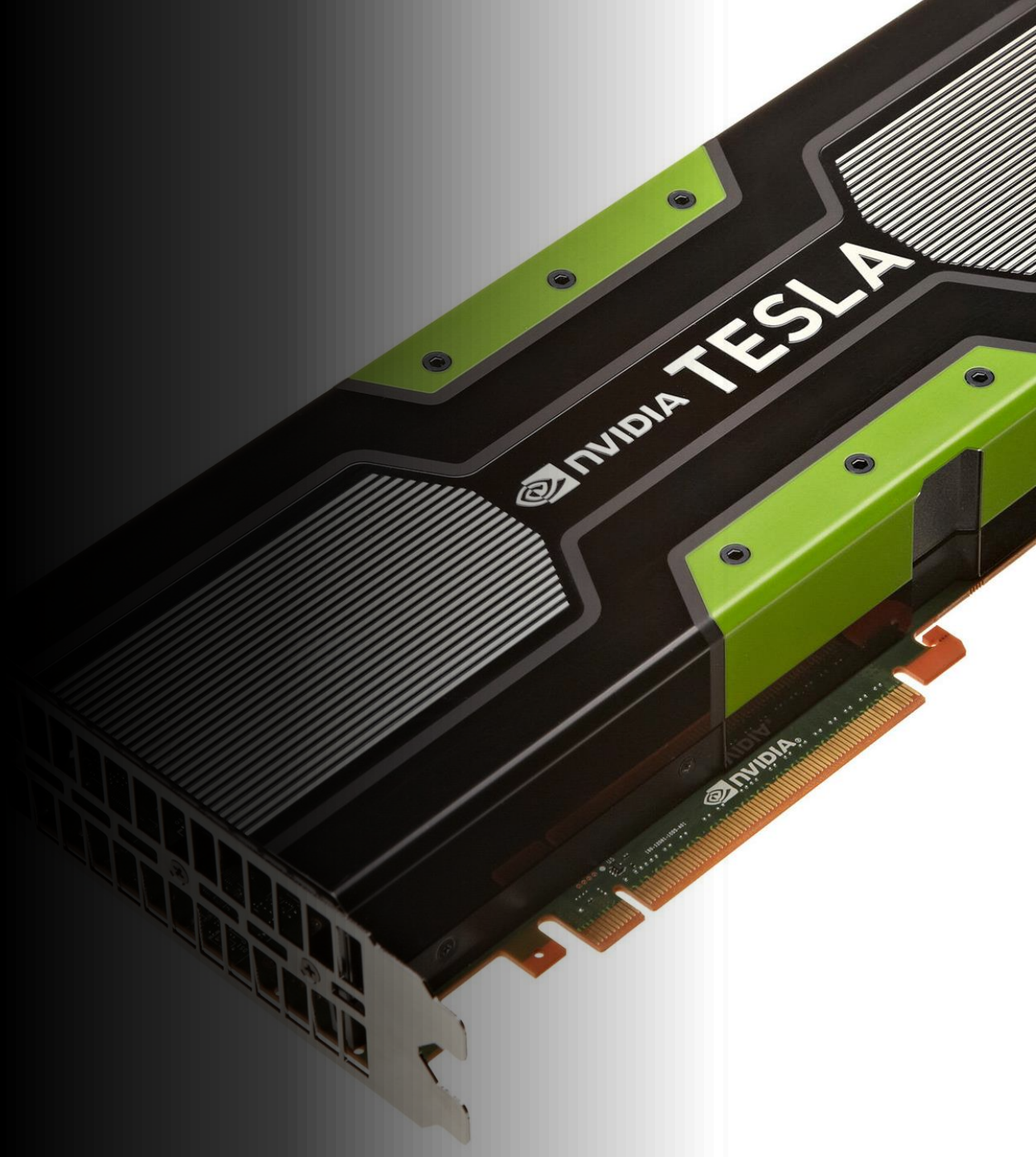
Hans Moravec (1988): *“It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and* ***difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility****.”*

A teenager can learn how to drive in a few hours. We still have no truly self-driving car.

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

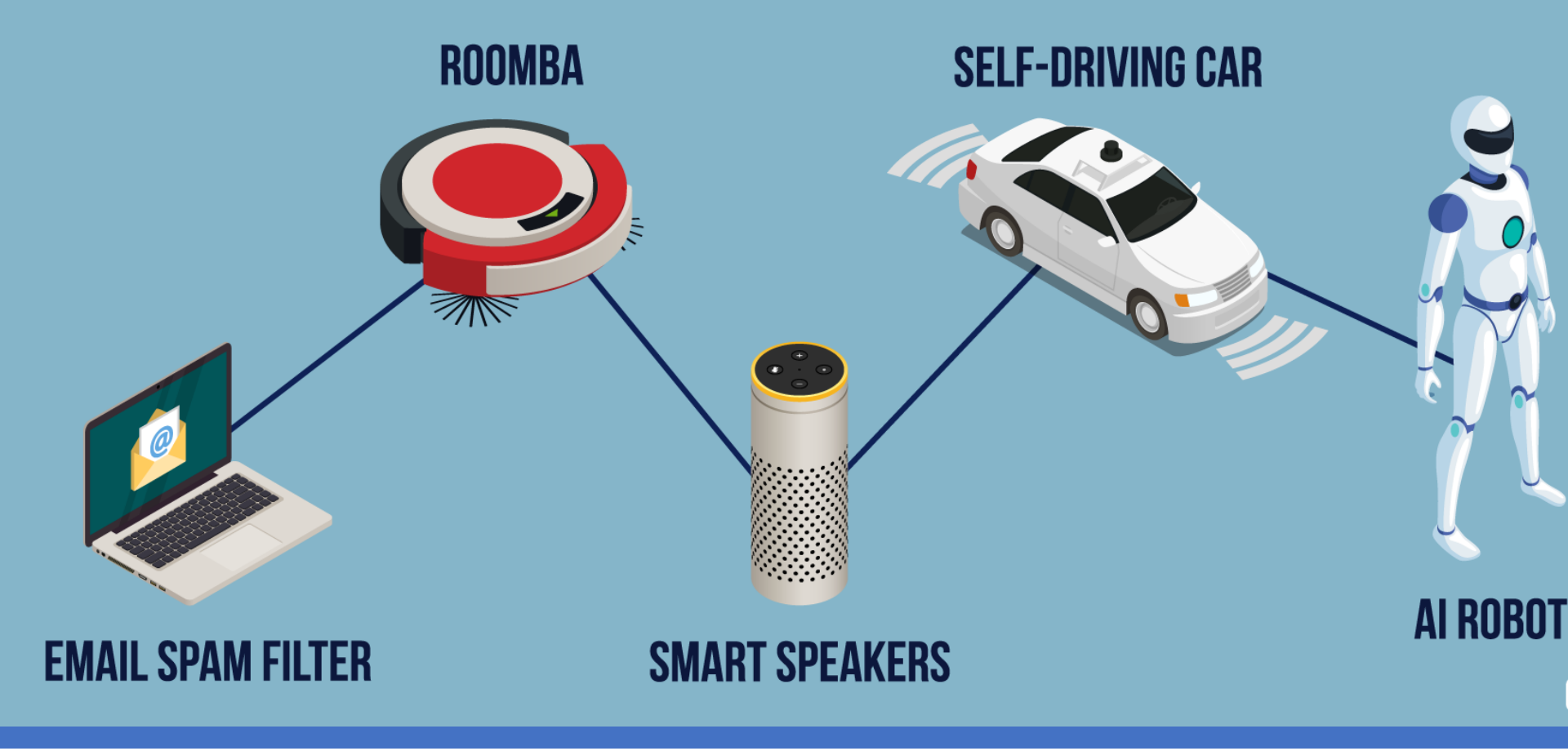
What accounts for recent successes in AI?

- 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).



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, massively parallel computers essentially use brute force search to beat grand masters
- Learning once thought uniquely human
  - Ada Lovelace (1842): “The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform.”
  - Now machine learning is a well-developed discipline
- Similar picture with animal intelligence… “Even a monkey can do this!”

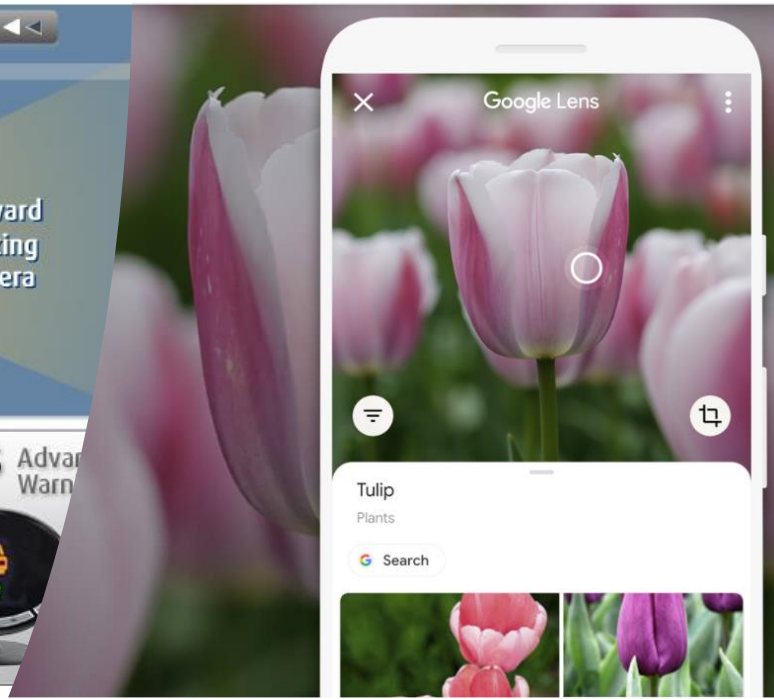

ROOMBA
SELF-DRIVING CAR
EMAIL SPAM FILTER
SMART SPEAKERS
AI ROBOT

AI Today

Vision and Image Processing

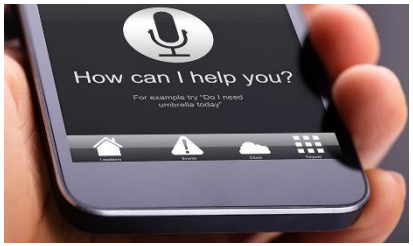
- **OCR**: read license plates, handwriting recognition (e.g., mail sorting).
- **Face detection**: now standard for smart phone cameras.
- **Vehicle safety systems**
- **Visual search**
- **Image generation**

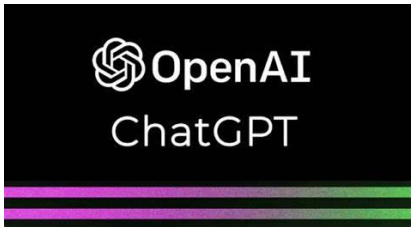
All these technologies can now operate now at superhuman performance.



Natural Language Processing

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

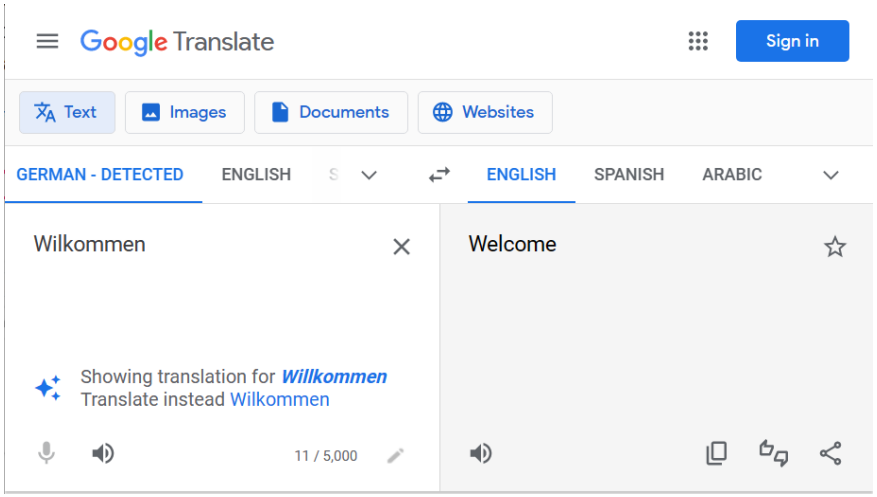




All these technologies can operate now with close to or even superhuman performance.

Humans use language to reason. Does that mean AI that can create good language can reason?

Language understanding is still elusive!



Robotics

- Mars rovers
- Autonomous vehicles
  - DARPA Grand Challenge
  - Google self-driving cars
- Autonomous helicopters and drones
- Robot soccer
  - RoboCup
- Personal robotics
  - Humanoid robots
  - Robotic pets
  - Personal assistants?

Question Answering: IBM Watson

- Listens to spoken language.
- Speaks.
- **Finds questions to factual answers**.

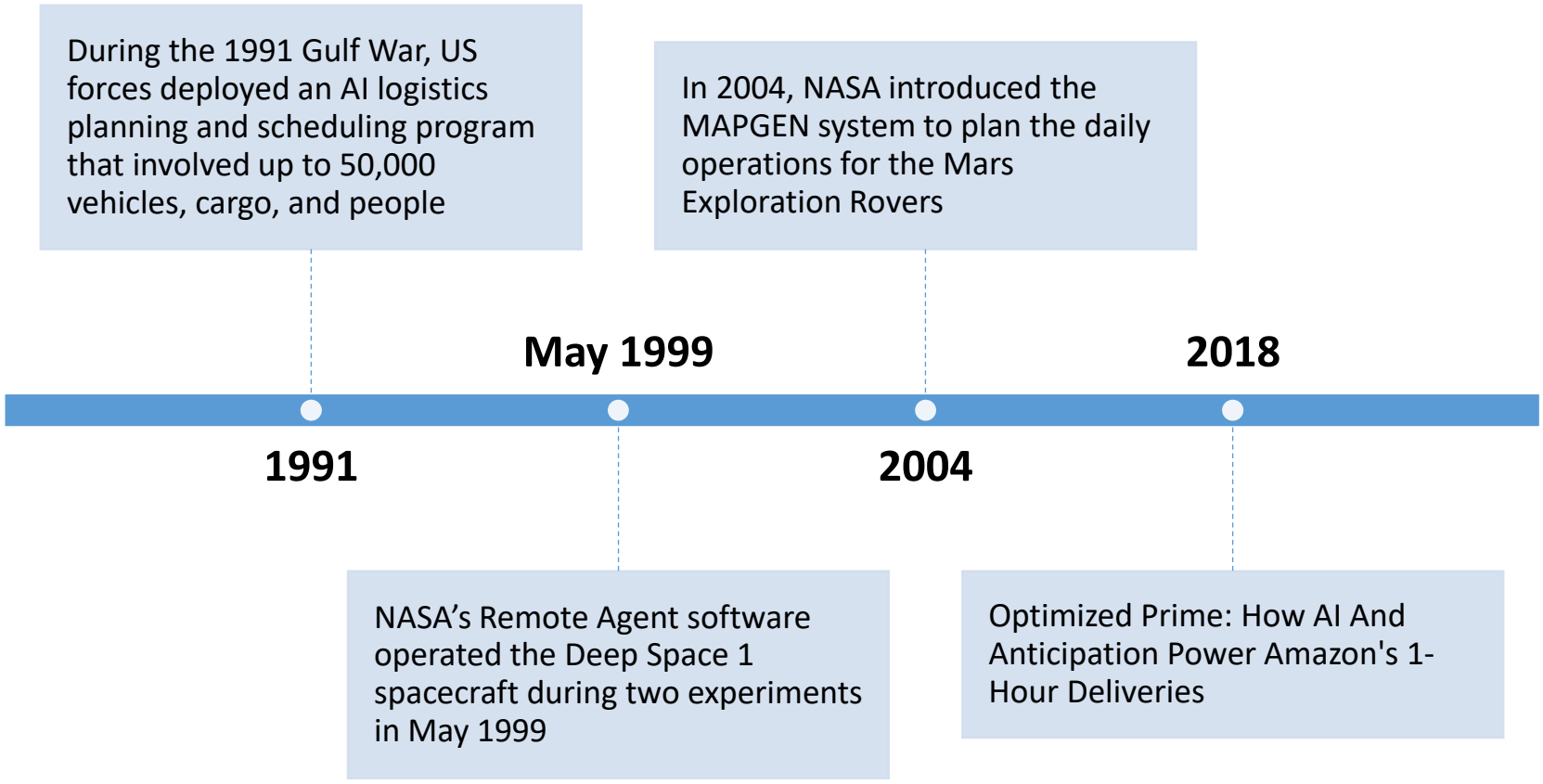
- http://www.research.ibm.com/deepqa/
- NY Times article
- Trivia demo
- YouTube video
- IBM Watson wins on Jeopardy (February 2011)

Math, Games and Puzzles

- 1996: A computer program written by researchers at Argonne National Laboratory **proved a mathematical conjecture** (Robbins conjecture) unsolved for decades
  - NY Times story: "[The proof] would have been called creative if a human had thought of it"
- 1996/97: IBM's Deep Blue defeated the reigning world chess champion Garry Kasparov in 1997
  - **1996: Kasparov Beats Deep Blue**
    "I could feel --- I could smell --- a new kind of intelligence across the table."
  - **1997: Deep Blue Beats Kasparov**
    "Deep Blue hasn't proven anything."
- 2007**: Checkers** was "solved" --- a computer system that never loses was developed. Science article
- 2017+: AlphaZero learns **chess, shogi and go** by playing itself. Science article
- 2019: MuZero learns to play **Atari computer games**.

AI exhibits superhuman performance on almost all games.

Logistics, Scheduling and Planning



Information agents

- Search engines
- Chat Agents / Automated helpdesks
- Recommendation systems
- Spam filtering
- Medical diagnosis systems
- Fraud detection
- Automated stock trading

Self-driving Cars and Safety

SAE Automation Levels

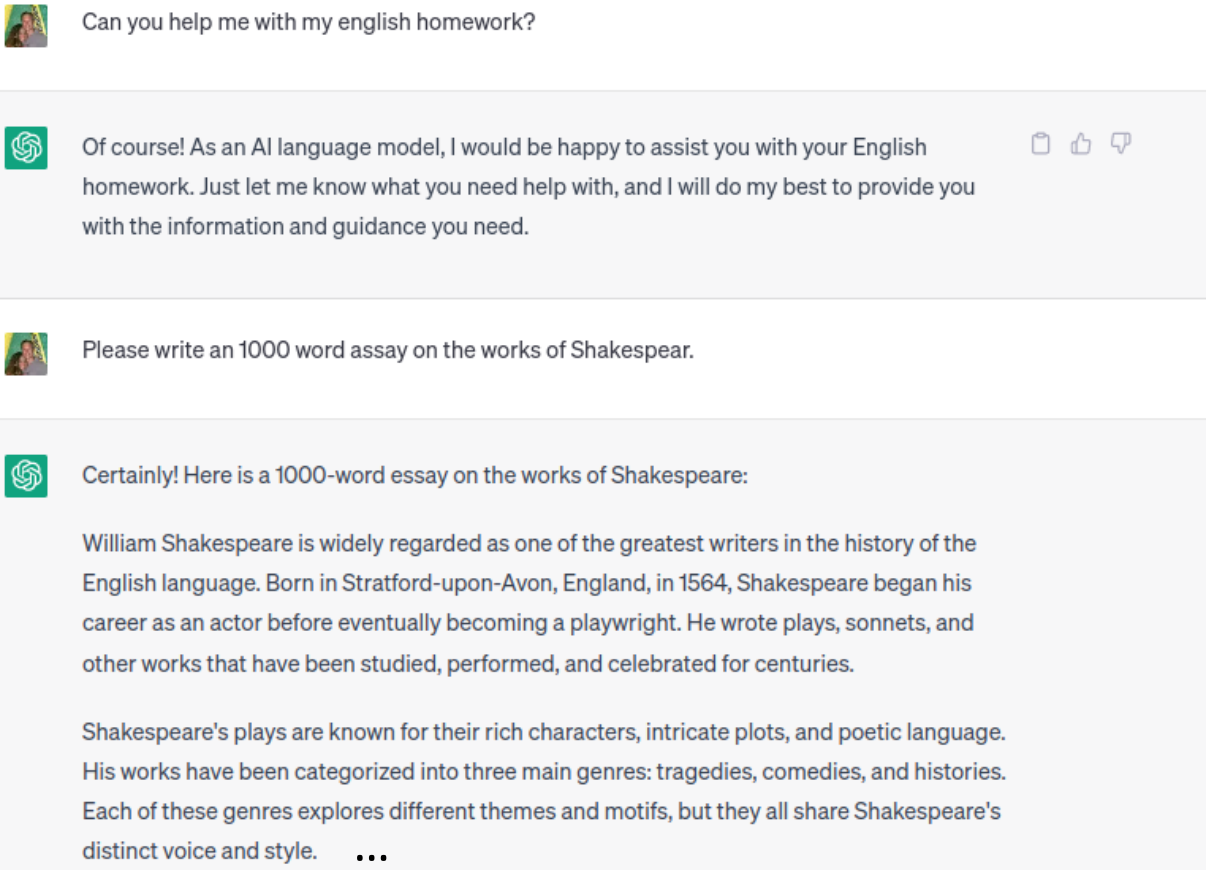
- Level 1 - Driver Assistance (“hands on”)
- Level 2 - Partial Automation (“hands off”)
- Level 3 - Conditional Automation
- Level 4 - High Automation
- Level 5 - Full Automation (“steering wheel optional”)

Components

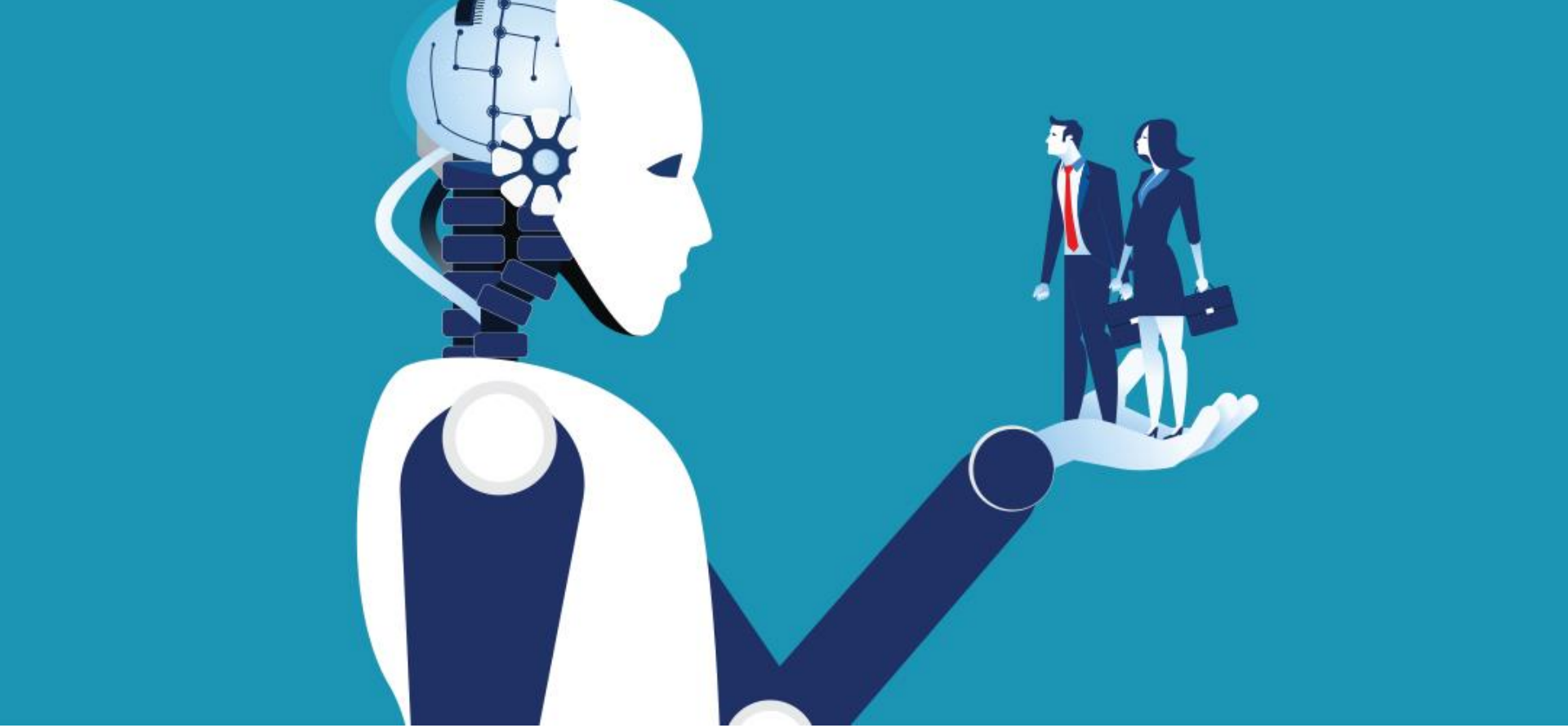
- Sensing
- Maps
- Path planning
- Controlling the vehicle

Why is this so hard?

Large Language Models (LLMs)



Source: https://chat.openai.com/



AI Ethics & Safety

A new Frontier for Fairness and Freedom

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
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.

Australia: AI Ethics Framework for Industry
A set of **voluntary AI Ethics Principles** (2019)

Core principles for AI	
1. Generates net-benefits. The AI system must generate benefits for people that are greater than the costs.	**2. Do no harm.** Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes.
3. Regulatory and legal compliance. The AI system must comply with all relevant international, Australian Local, State/Territory and Federal government obligations, regulations and laws.	**4. Privacy protection.** Any system, including AI systems, must ensure people's private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm.
5. Fairness. The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly.	**6. Transparency & Explainability.** People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.
7. Contestability. When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.	**8. Accountability.** People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.

https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

In the US

Bill Introduced in 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

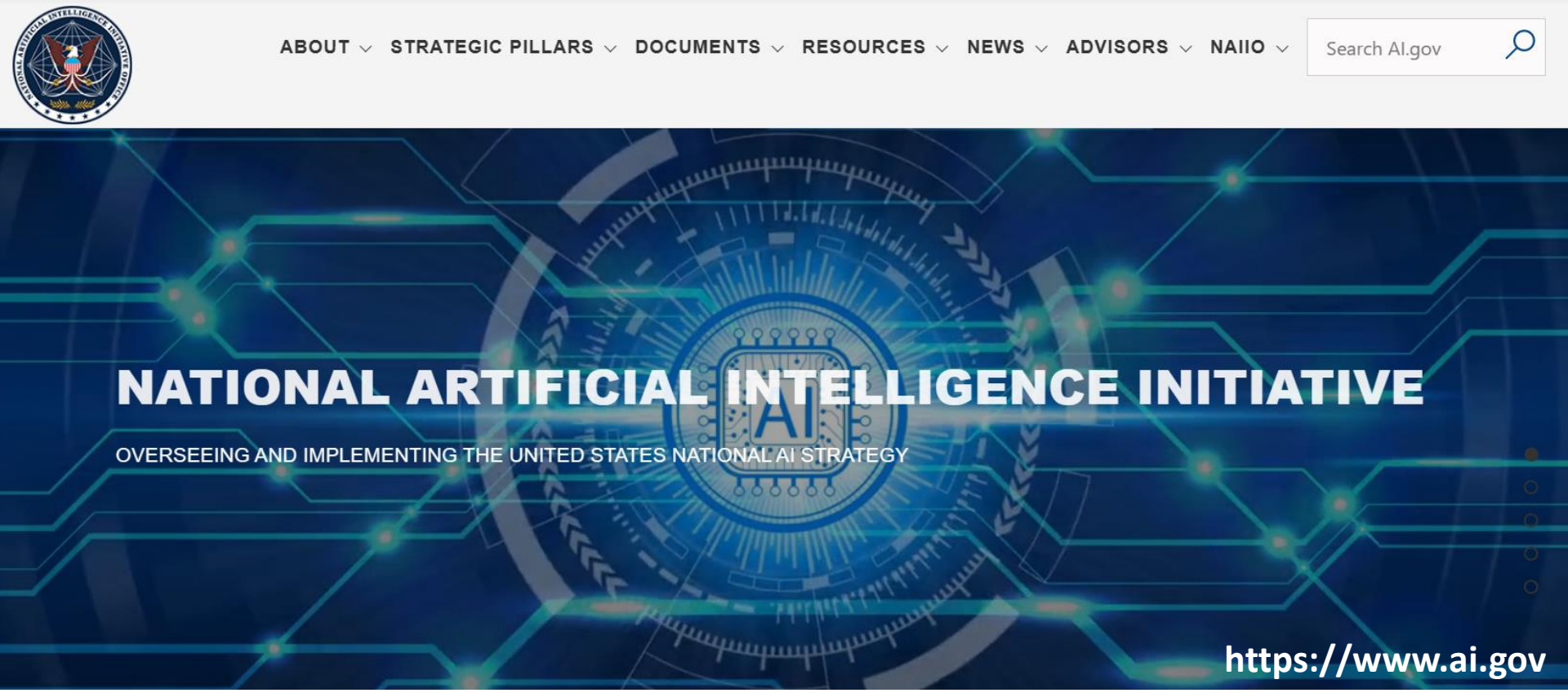
(D) the measures the covered entity will employ to minimize the risks described in subparagraph (C), including technological and physical safeguards.

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



US National AI Initiative Act of 2020

"The National AI Initiative Act of 2020 became law on January 1, 2021, providing for a coordinated program across the entire Federal government to **accelerate AI research** and application for the Nation's economic prosperity and national security. The mission of the National AI Initiative is to ensure continued U.S. leadership in AI research and development, lead the world in the development and use of trustworthy AI in the public and private sectors, and prepare the present and future U.S. workforce for the integration of AI systems across all sectors of the economy and society."

May also in the future provide regulations

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

Algorithmic Bias and Fairness

"**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.
- For example: Evaluate job applicants for a job which is historically almost exclusively held by males.

Technical bias

- Limitations of a program or computational power.
- For example: instead of a random sample, the program uses the first n data points.

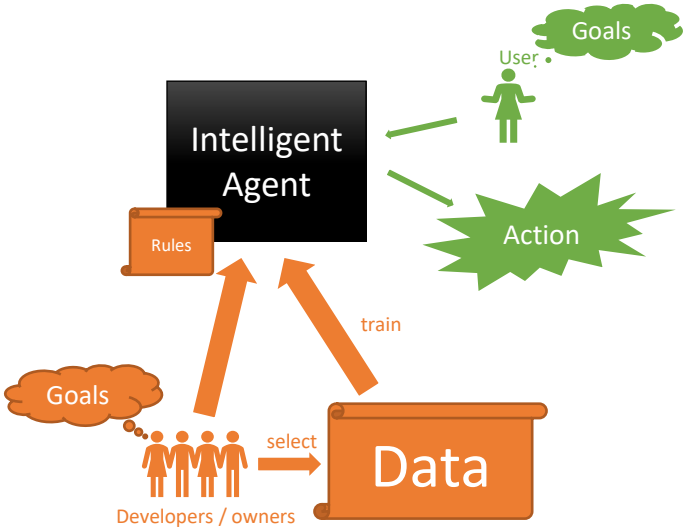
Emergent bias

- Use of algorithms for new data without checking for bias (e.g., existing correlations in the data).
- Use of an algorithm for an unanticipated application.

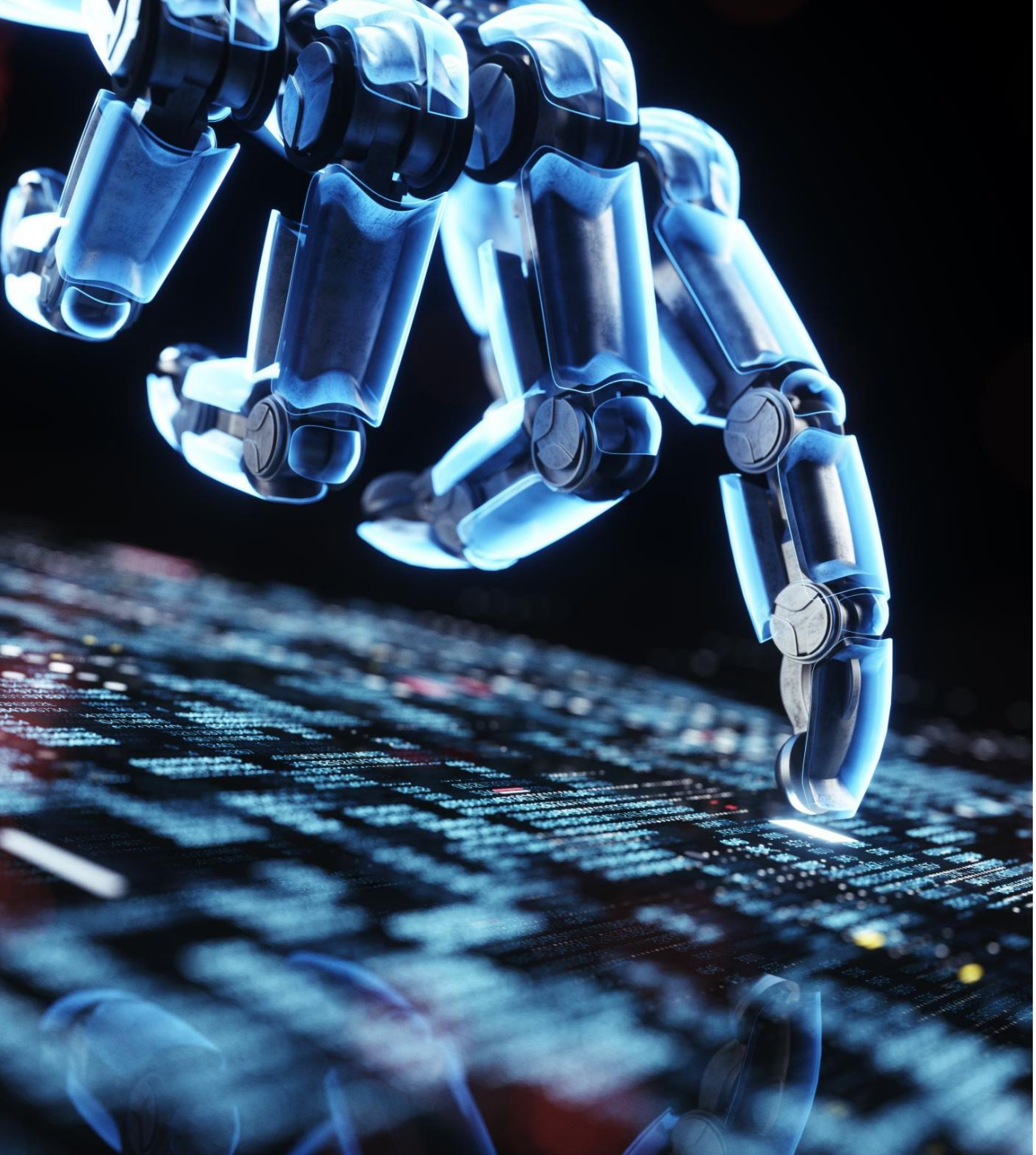
AI Safety

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

- Robustness: Black swan vs. adversarial robustness
- Monitoring AI
- What about liability?
- Goal/reward alignment
- Reward hacking
- AGI and instrumental convergence



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**.

- **This course will introduce simple techniques to create intelligent agents.**