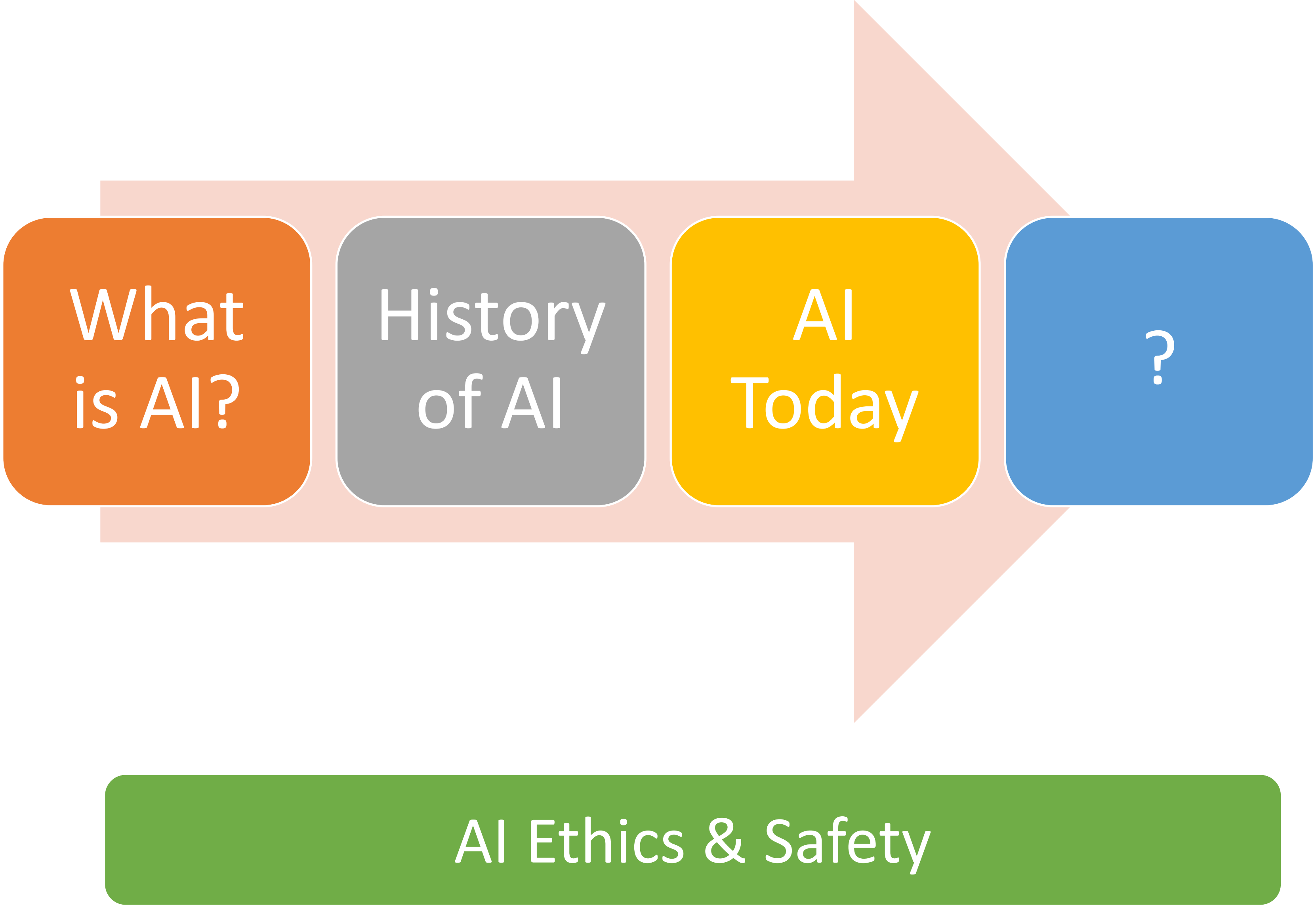
CS 5/7320 Artificial Intelligence

Introduction

AIMA Chapters 1 + 27

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



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.”

- We call the machine an **agent** or **intelligent agent.**

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

How do 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.

- Requires scientific theories of how the brain works.

Note: The brain does not work like artificial neural networks from ML!

How to understand cognition as a computational process?

- Introspection: try to think about how we think.
- Predict the behavior of human subjects.
- Image the brain, examine neurological data

AI consciousness

- What does it mean that a machine is conscient/sentient?
- How can we tell?

(What do we do?)

Cognitive Sciences

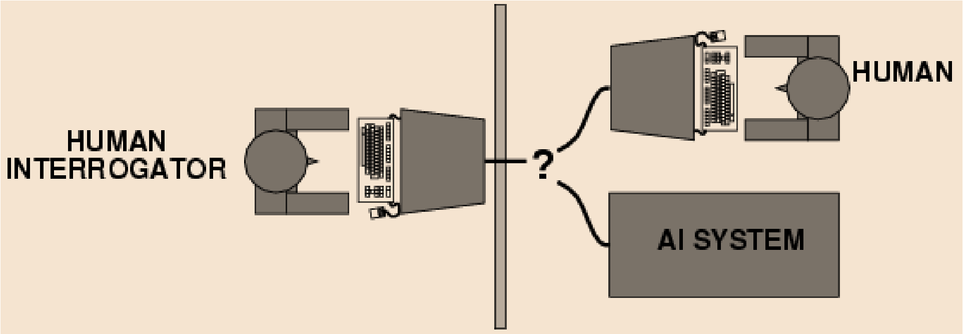
Think like a human?

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

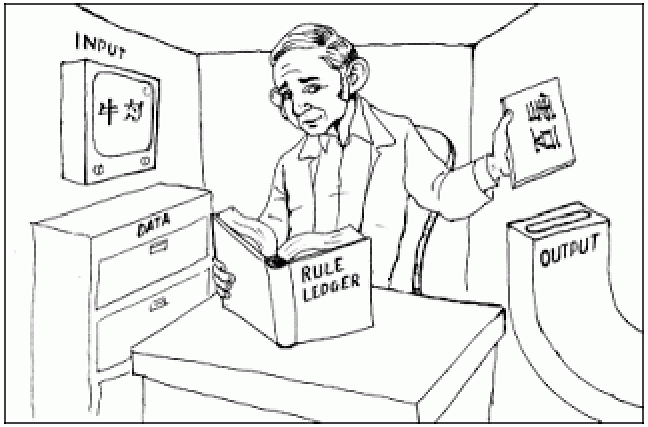
What are some potential problems with the Turing Test?

- 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?

- **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 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.

What about the logical implication

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

Should it rather be

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

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 rules.
  - **Practicality**: can be adapted to many real-world problems.
  - **Well established**: solvers, 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.

What type of AI do we cover in this course?

Create a **narrow AI agent** that can

~~think like a human?~~

~~act like a human?~~

~~think rationally?~~

act rationally.

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

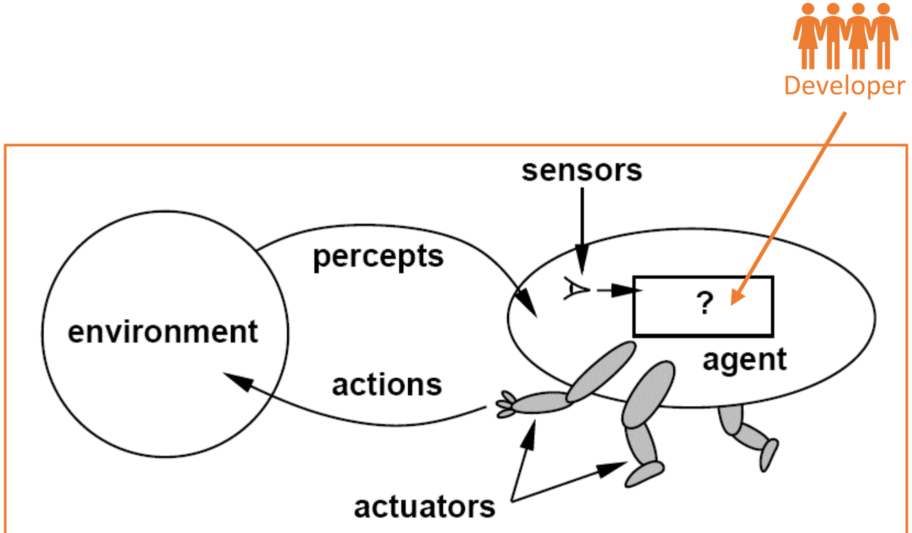
What are the Components of an Intelligent Agent?

Intelligent agents need to

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

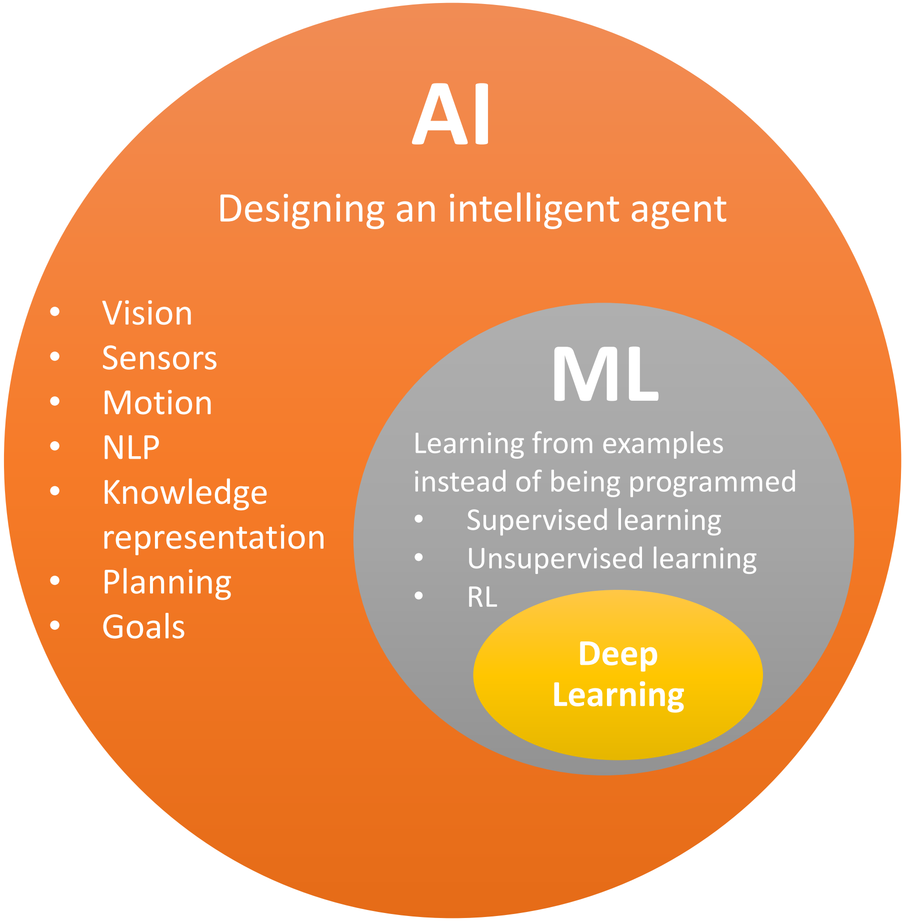
Optional

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



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

Machine Learning vs. Artificial Intelligence



AI

Designing an intelligent agent

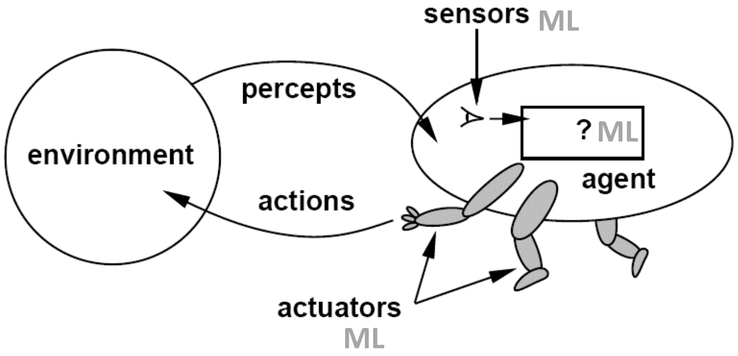
- Vision
- Sensors
- Motion
- NLP
- Knowledge representation
- Planning
- Goals

ML

Learning from examples instead of being programmed

- Supervised learning
- Unsupervised learning
- RL

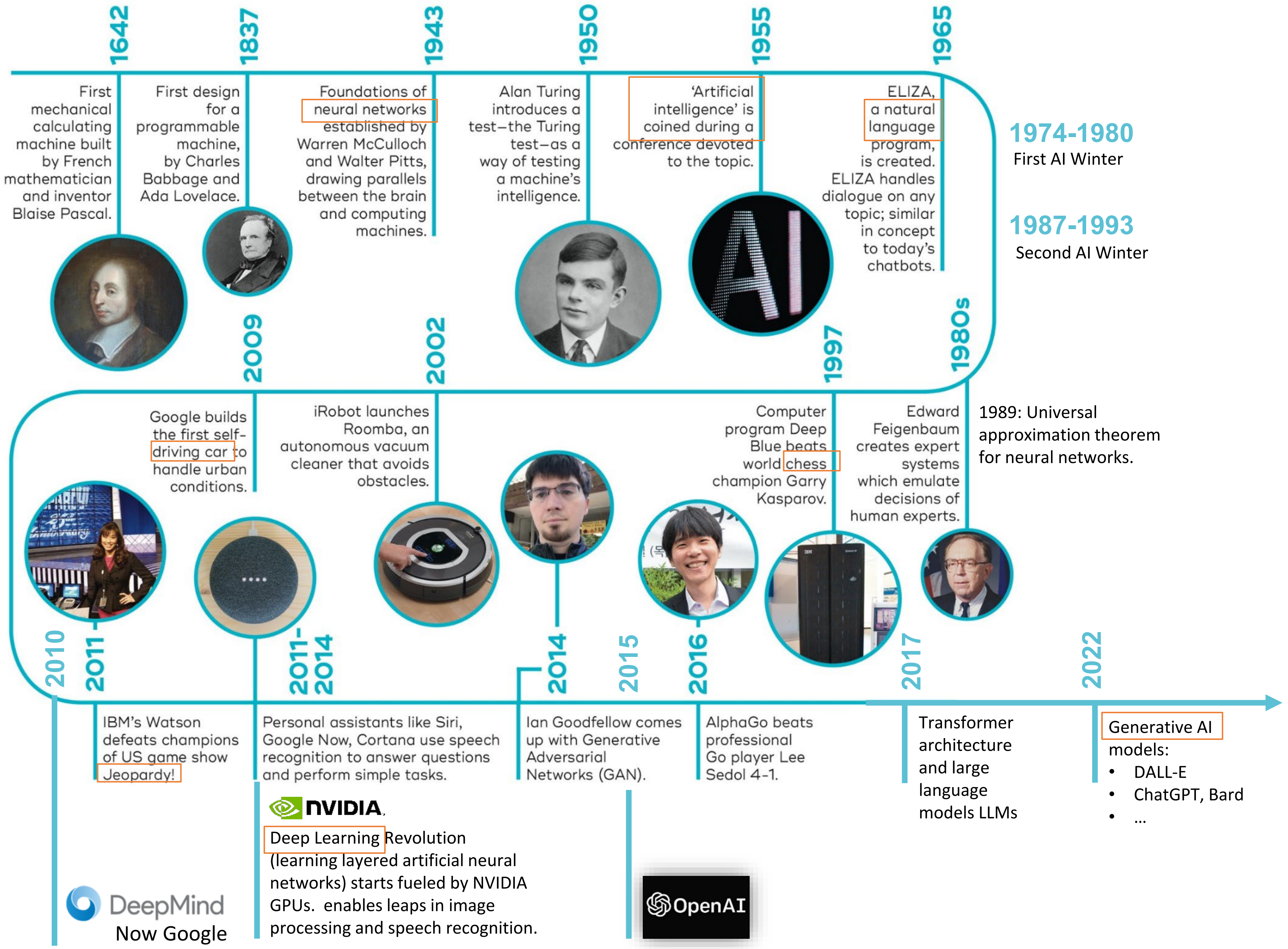
Deep Learning



Agent interacting with the environment

[Artificial Intelligence: A Modern Approach, Editions 1-3]

The History of AI



1642
First mechanical calculating machine built by French mathematician and inventor Blaise Pascal.

1837
First design for a programmable machine, by Charles Babbage and Ada Lovelace.

1943
Foundations of neural networks established by Warren McCulloch and Walter Pitts, drawing parallels between the brain and computing machines.

1950
Alan Turing introduces a test–the Turing test–as a way of testing a machine's intelligence.

1955
'Artificial intelligence' is coined during a conference devoted to the topic.

1965
ELIZA, a natural language program, is created. ELIZA handles dialogue on any topic; similar in concept to today's chatbots.

1974-1980
First AI Winter

1987-1993
Second AI Winter

1980s
Edward Feigenbaum creates expert systems which emulate decisions of human experts.

1989: Universal approximation theorem for neural networks.

1997
Computer program Deep Blue beats world chess champion Garry Kasparov.

2002
iRobot launches Roomba, an autonomous vacuum cleaner that avoids obstacles.

2009
Google builds the first self-driving car to handle urban conditions.

2010
DeepMind
Now Google

2011
IBM's Watson defeats champions of US game show Jeopardy!

2011-2014
Personal assistants like Siri, Google Now, Cortana use speech recognition to answer questions and perform simple tasks.

NVIDIA.
Deep Learning Revolution (learning layered artificial neural networks) starts fueled by NVIDIA GPUs. enables leaps in image processing and speech recognition.

2014
Ian Goodfellow comes up with Generative Adversarial Networks (GAN).

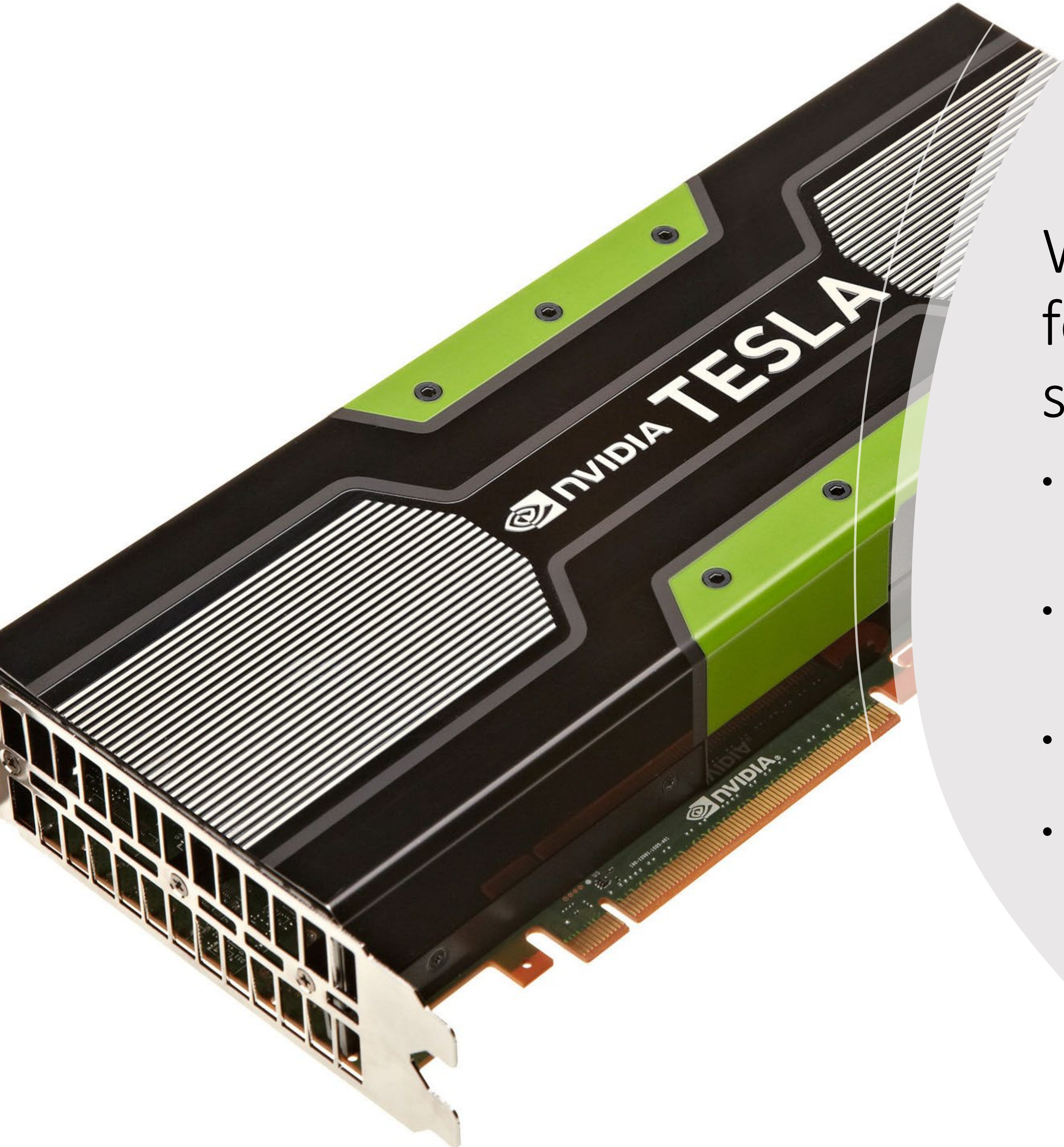
2015
OpenAI

2016
AlphaGo beats professional Go player Lee Sedol 4-1.

2017
Transformer architecture and large language models LLMs

2022
Generative AI models:
- DALL-E
- ChatGPT, Bard
- …

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”

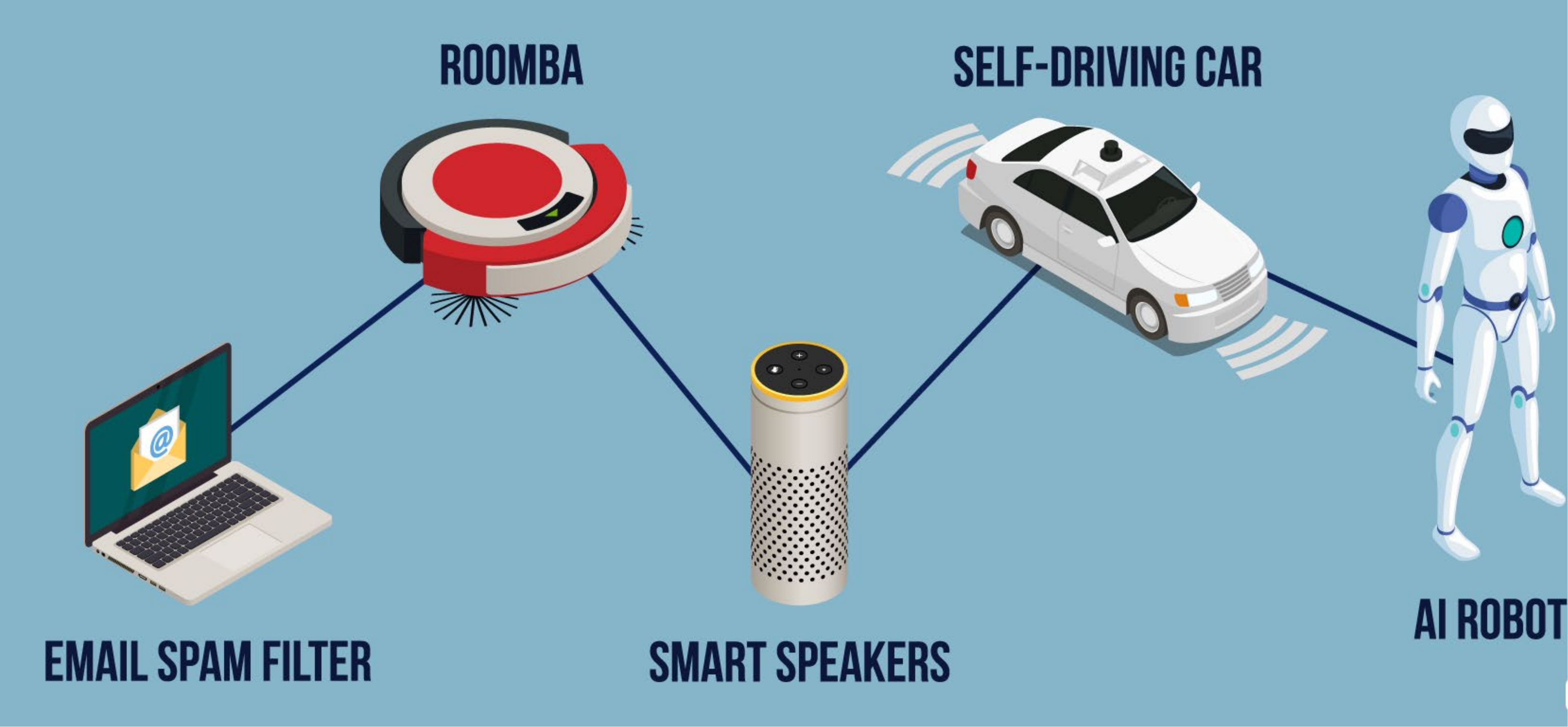
Hans Moravec (1988): *“It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and* ***difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility****.”*

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
- “Even a monkey can do this!”

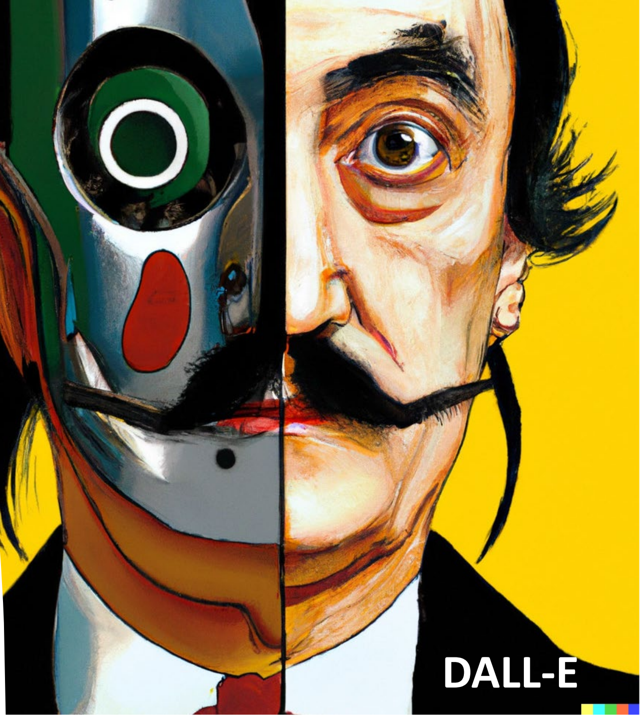

ROOMBA
SELF-DRIVING CAR
AI ROBOT
EMAIL SPAM FILTER
SMART SPEAKERS

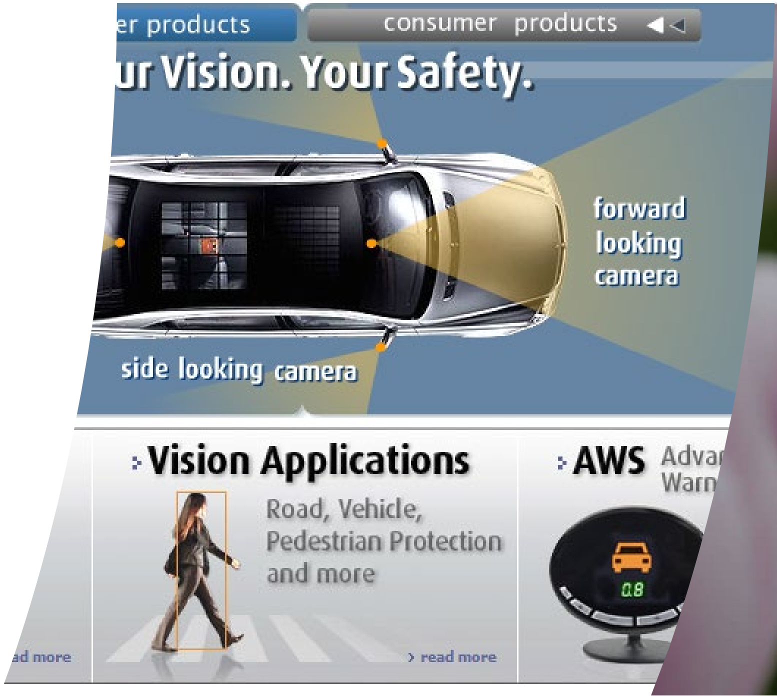
AI Today

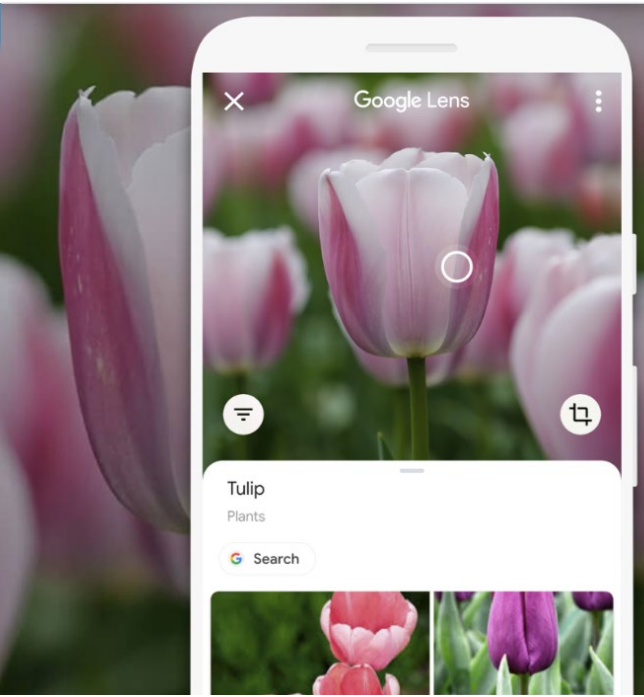
Vision and Image Processing

- **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.

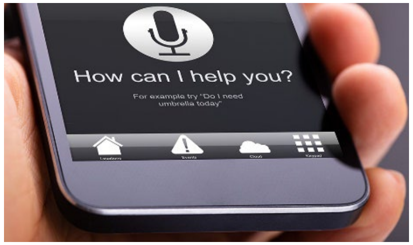


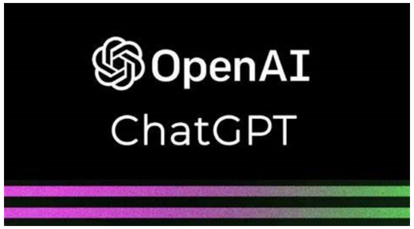




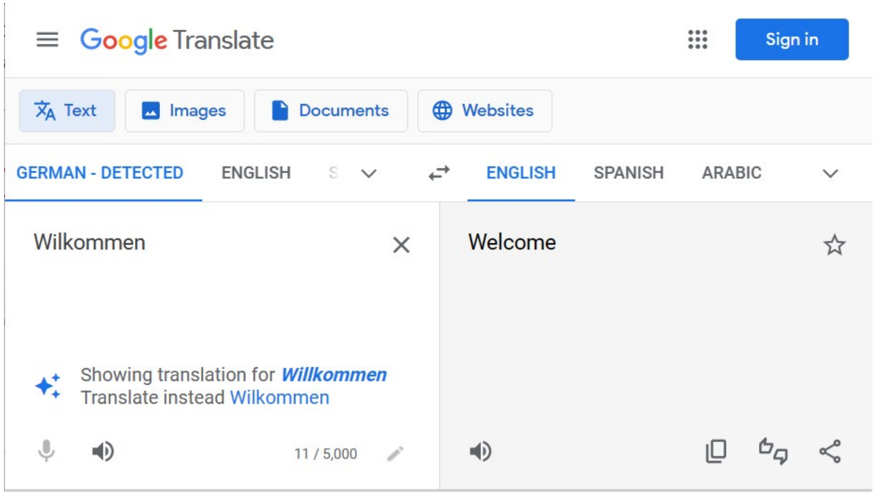
Natural Language Processing

- 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.



Google Translate

Sign in

Text | Images | Documents | Websites

GERMAN - DETECTED | ENGLISH | S

ENGLISH | SPANISH | ARABIC

Wilkommen

Welcome

Showing translation for *Willkommen*
Translate instead Wilkommen

11 / 5,000

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)

Self-driving Cars and Safety

SAE Automation Levels

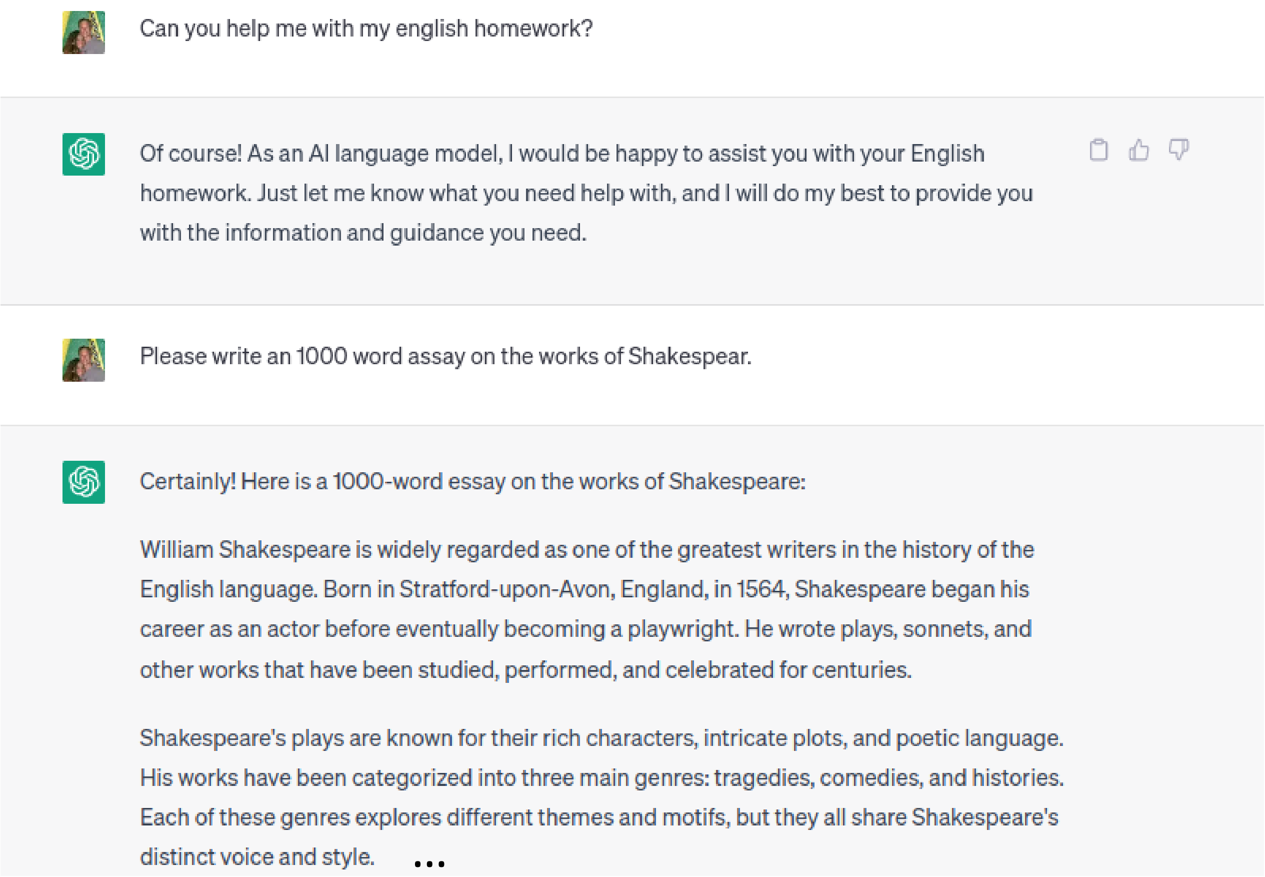
- Level 1 - Driver Assistance (“hands on”)
- Level 2 - Partial Automation (“hands off”)
- Level 3 - Conditional Automation
- Level 4 - High Automation
- Level 5 - Full Automation (“steering wheel optional”)

Components

- Sensing
- Maps
- Path planning
- Controlling the vehicle

Why is this so hard?

Large Language Models (LLMs)



Can you help me with my english homework?

Of course! As an AI language model, I would be happy to assist you with your English homework. Just let me know what you need help with, and I will do my best to provide you with the information and guidance you need.

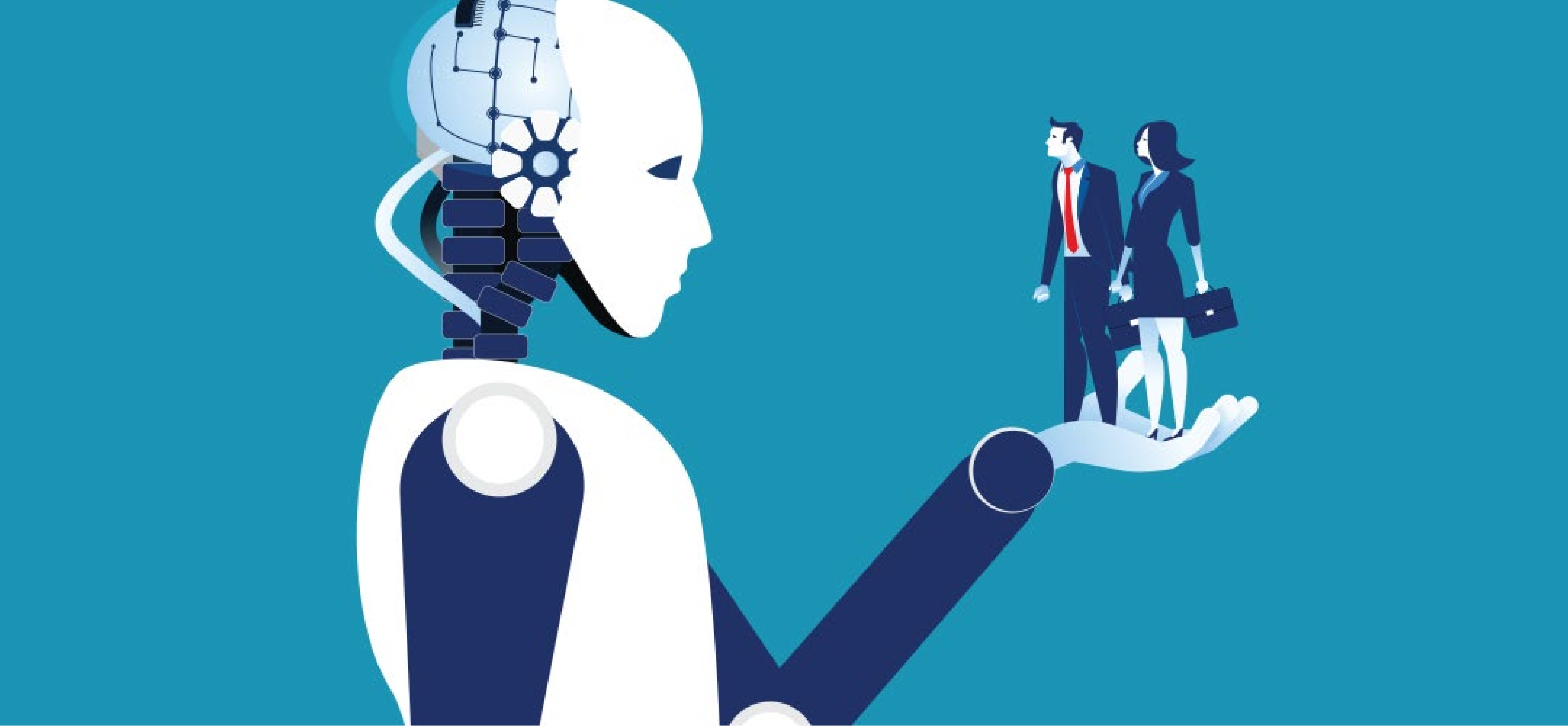
Please write an 1000 word assay on the works of Shakespear.

Certainly! Here is a 1000-word essay on the works of Shakespeare:

William Shakespeare is widely regarded as one of the greatest writers in the history of the English language. Born in Stratford-upon-Avon, England, in 1564, Shakespeare began his career as an actor before eventually becoming a playwright. He wrote plays, sonnets, and other works that have been studied, performed, and celebrated for centuries.

Shakespeare's plays are known for their rich characters, intricate plots, and poetic language. His works have been categorized into three main genres: tragedies, comedies, and histories. Each of these genres explores different themes and motifs, but they all share Shakespeare's distinct voice and style. …

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



AI Ethics & Safety

A new Frontier for Fairness and Freedom
AIMA Chapter 27

Commonly-Cited Principals

Use of AI by companies and organizations

- Ensure safety
- Limit harmful uses of AI
- Establish accountability: Liability?
- Avoid concentration of power: Winner-takes-All

Protect individuals

- Uphold human rights and values
- Ensure fairness: Equal opportunity/equal impact. Reflect diversity/inclusion
- Provide transparency: Explanations to build trust
- Respect privacy: Surveillance?
- Contemplate implications for employment: Income and purpose.

Governance

- Acknowledge legal/policy implications

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

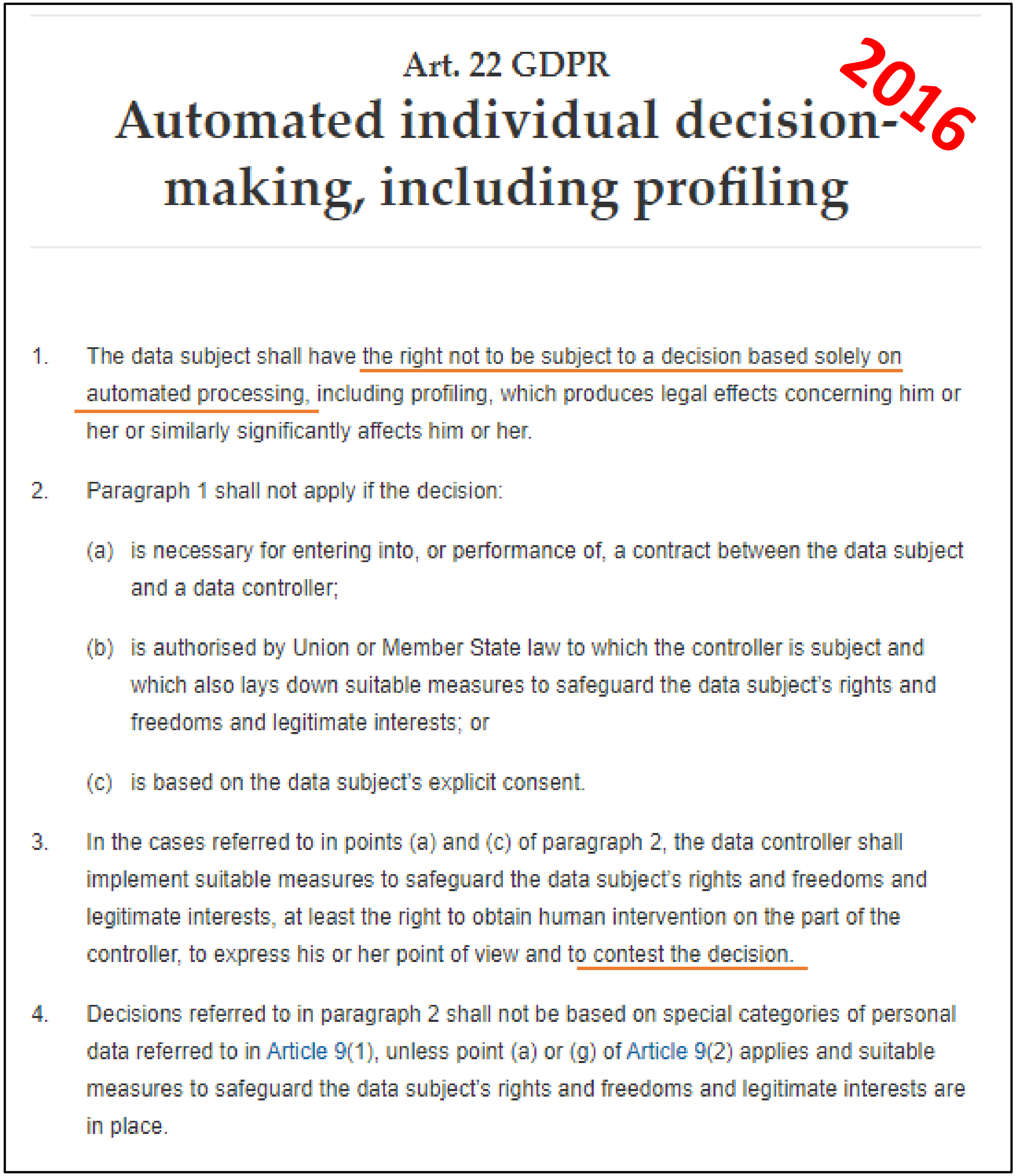
European Union

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



2016

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

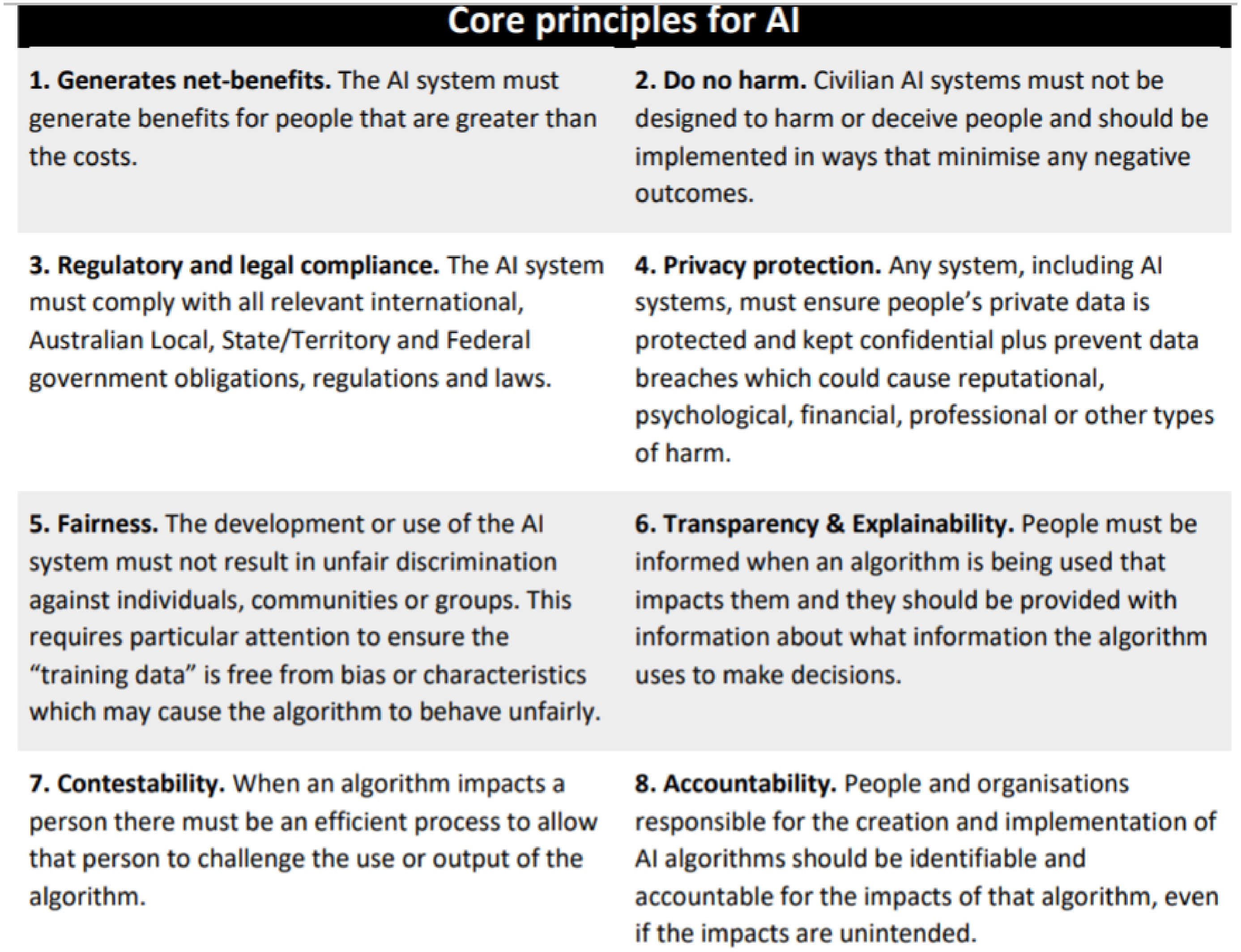
   (c) is based on the data subject's explicit consent.

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)

2019



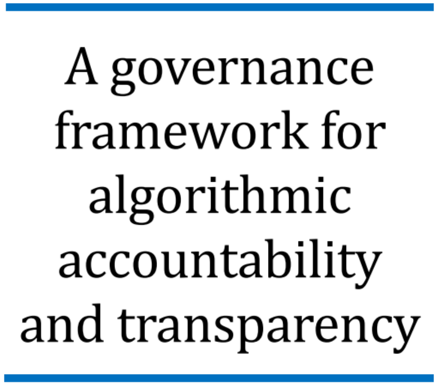
Core principles for AI

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.

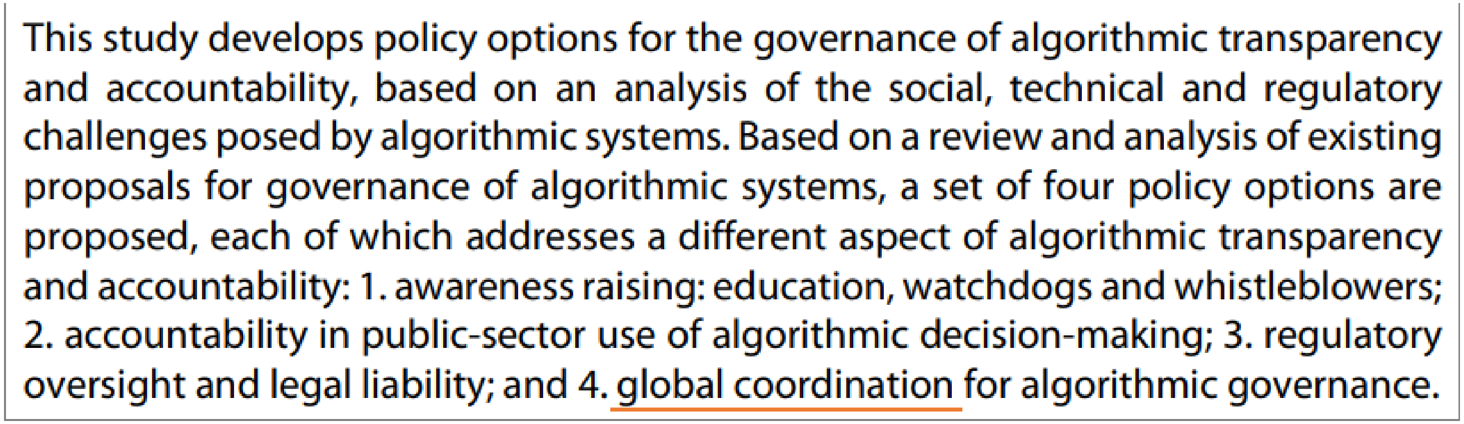
https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

European Union Study

2019

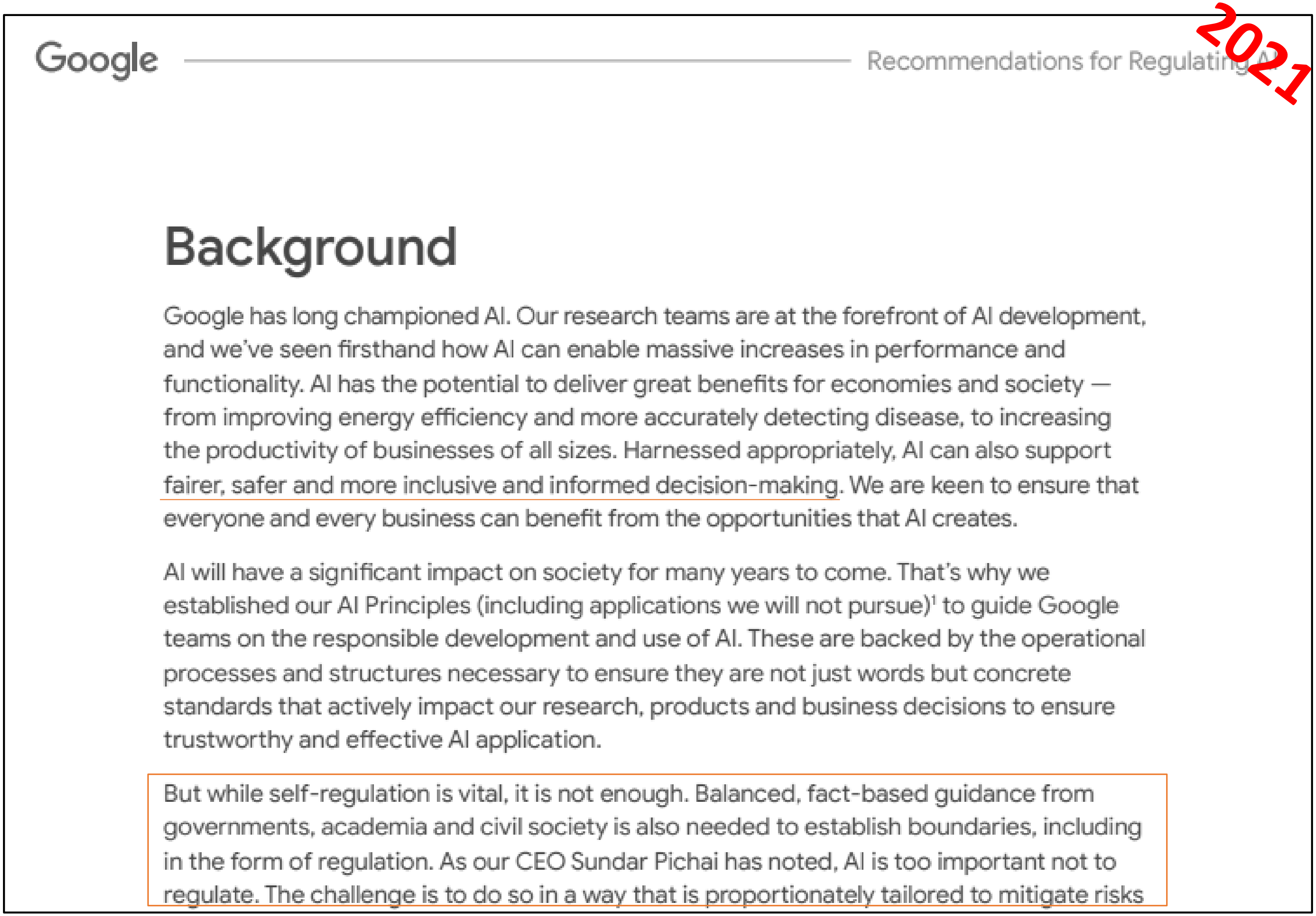


A governance framework for algorithmic accountability and transparency



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

Google Recommendations for Regulating AI

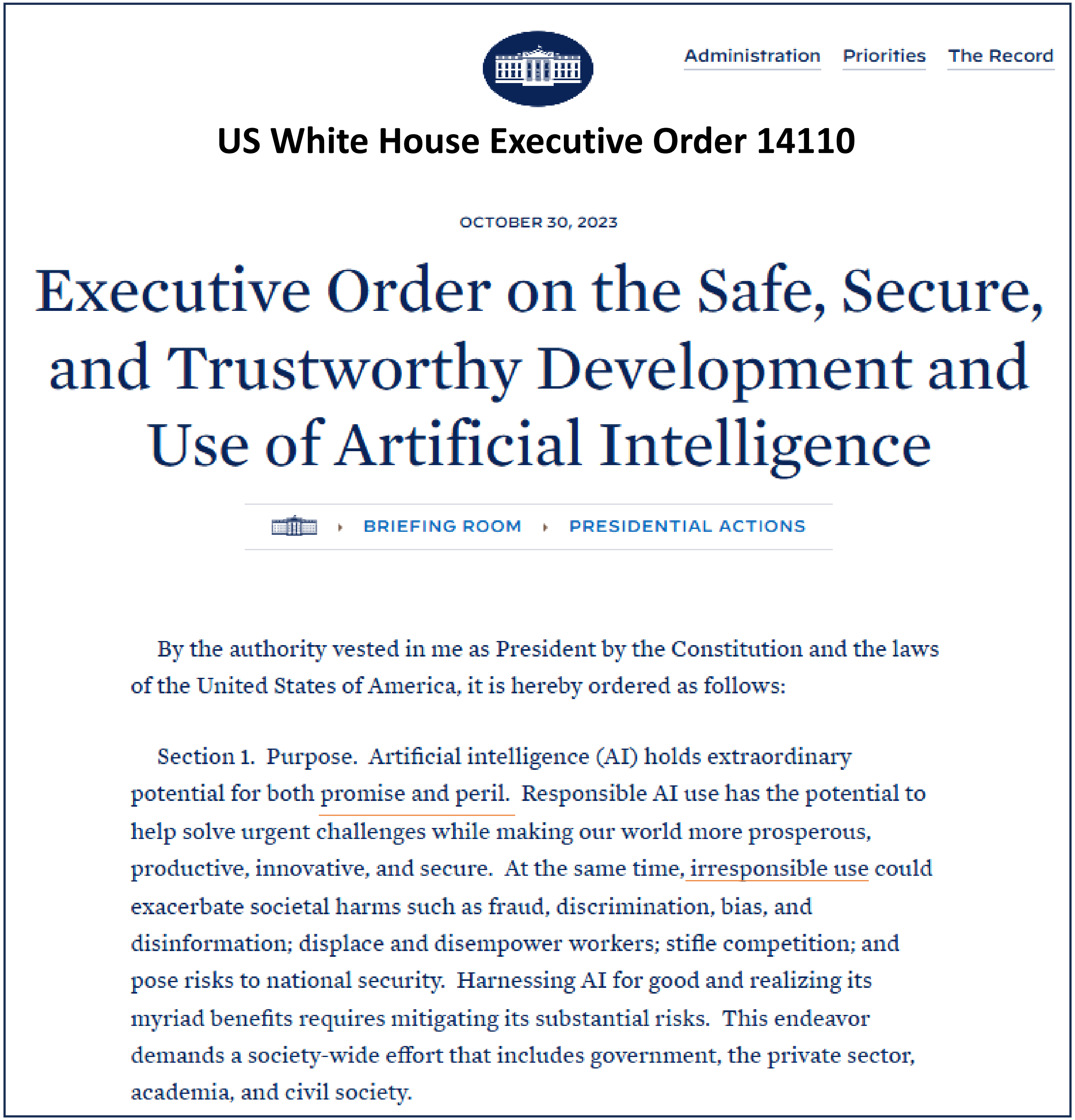
Background

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



US White House Executive Order 14110

OCTOBER 30, 2023

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

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.

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.

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

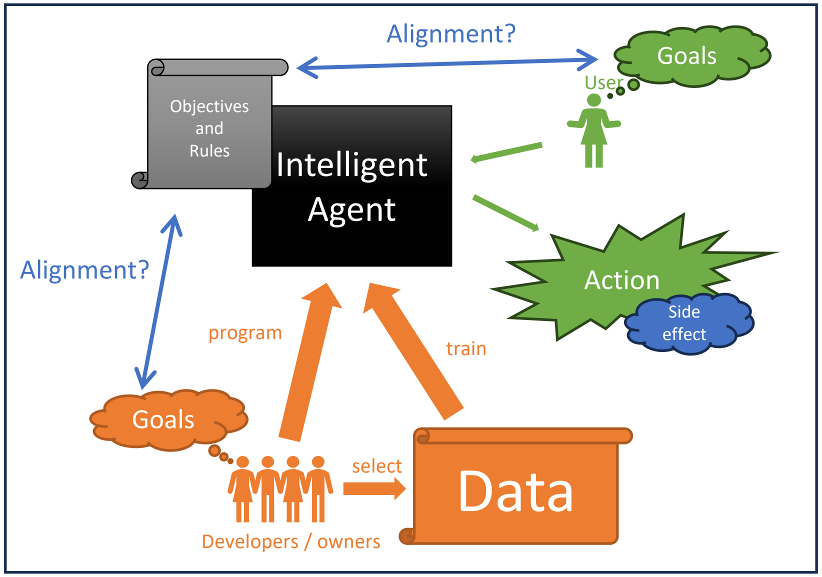
Use of algorithms for new data without checking for bias (e.g., existing correlations in the data).

Example: Use of an algorithm for an unanticipated application.

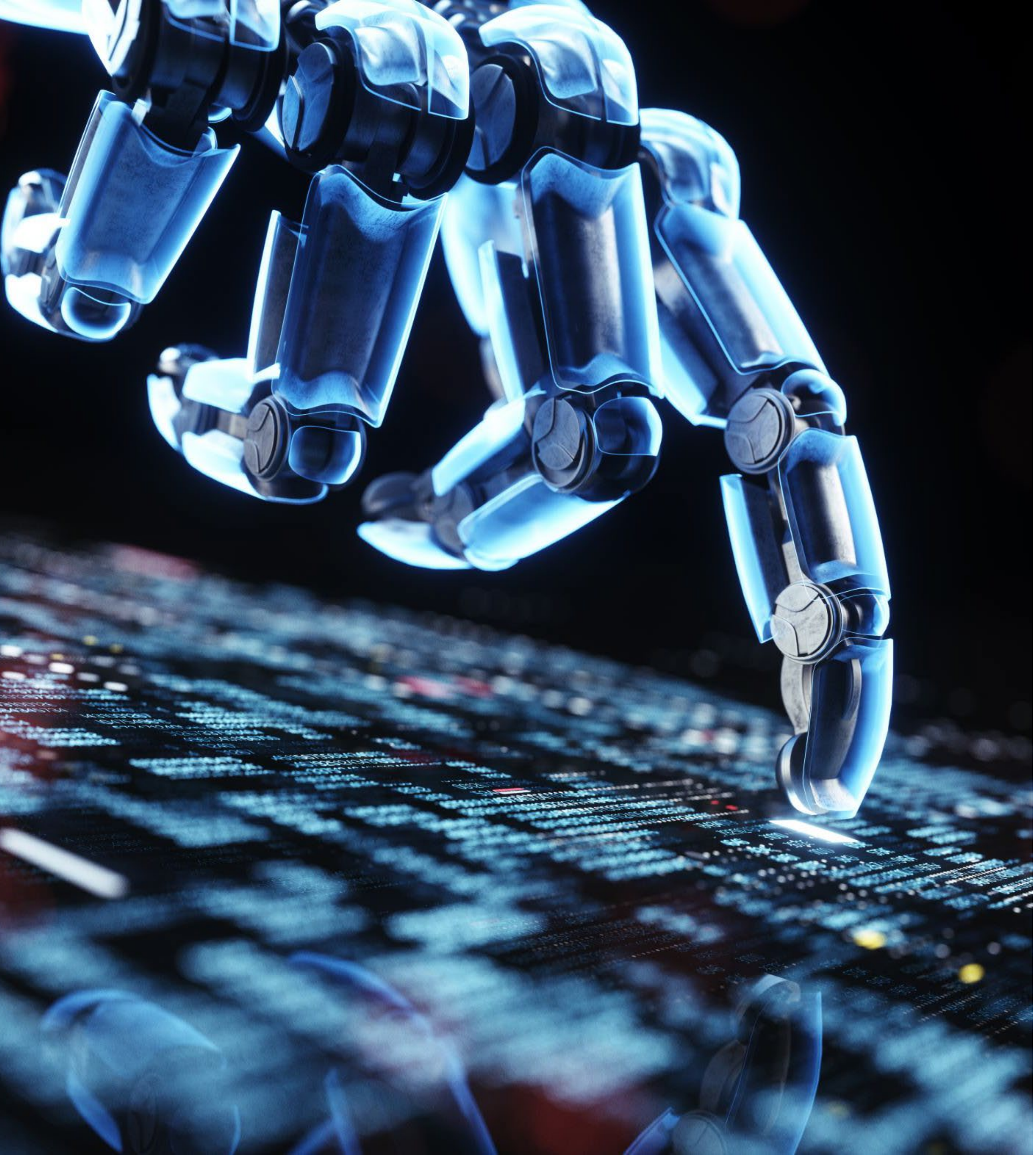
AI Safety

“Prevent accidents, misuse, or other harmful consequences of AI.”

- AI is an “optimizer!”
  - Goal/reward alignment: How do we specify a robust objective function?
  - Unintended side effects? AI needs to follow social norms.
  - Prevent reward hacking.
  - Instrumental convergence (common subgoals like the need for more power).
- AI Testing
- Monitoring AI
- Adversarial robustness



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.**