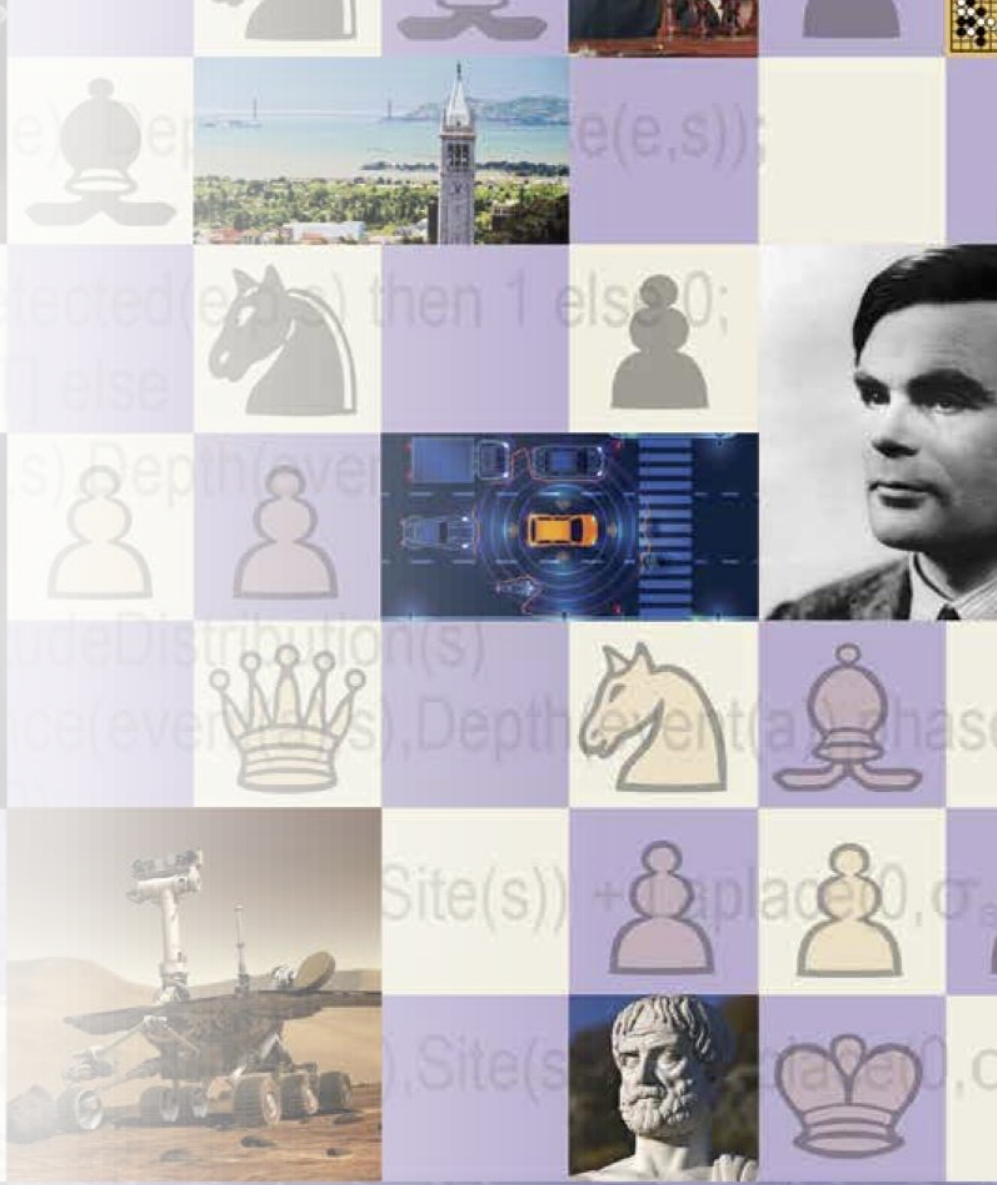
Discussion

CS 5/7320 Artificial Intelligence

Introduction

AIMA Chapter 1

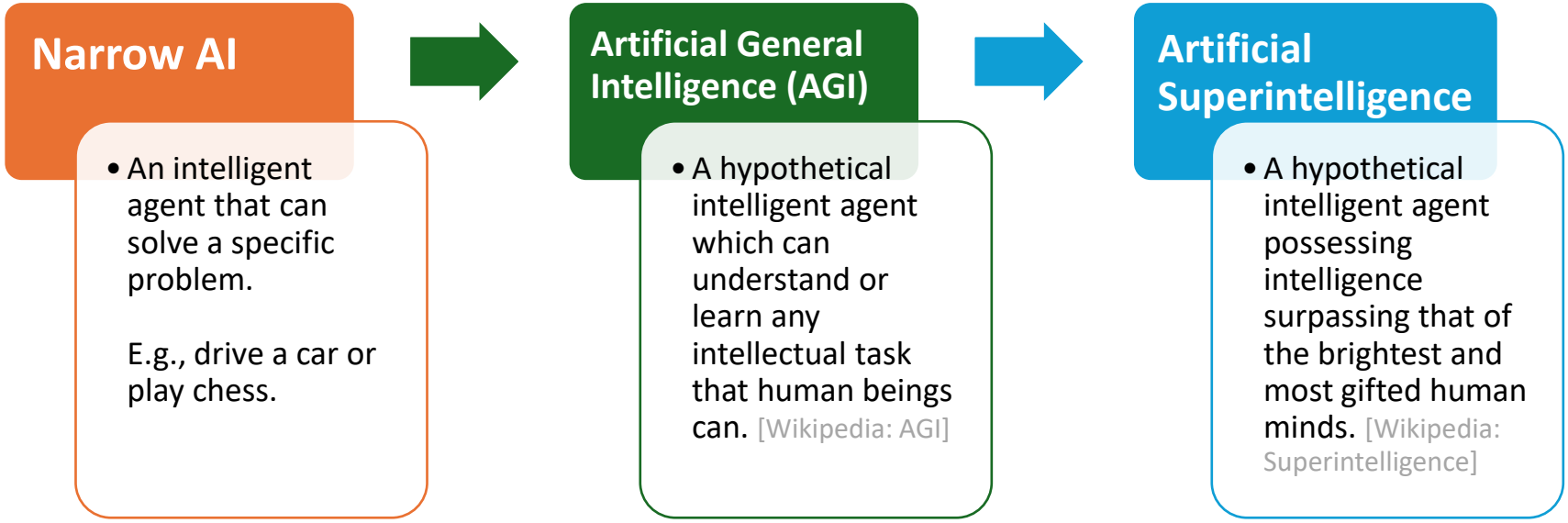
Slides by Michael Hahsler
with figures and cover art from
the AIMA textbook.



Module Review

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?

Components of an Intelligent Agent

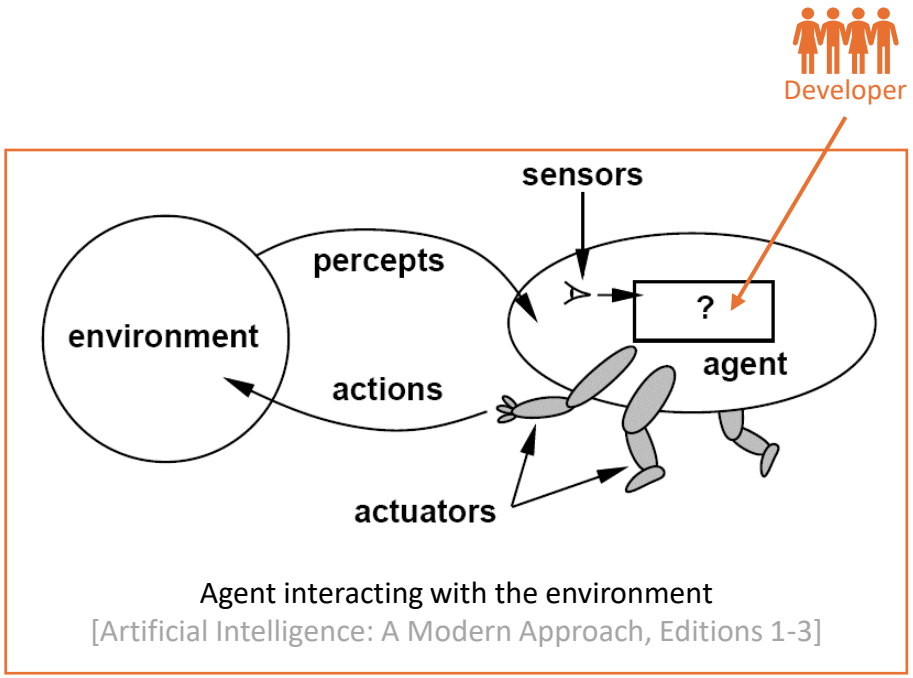
Intelligent agents **act rationally** in their environment.

They need to

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

Optional

- **Learn from experience** to improve performance**.** This uses **Machine Learning (ML)**

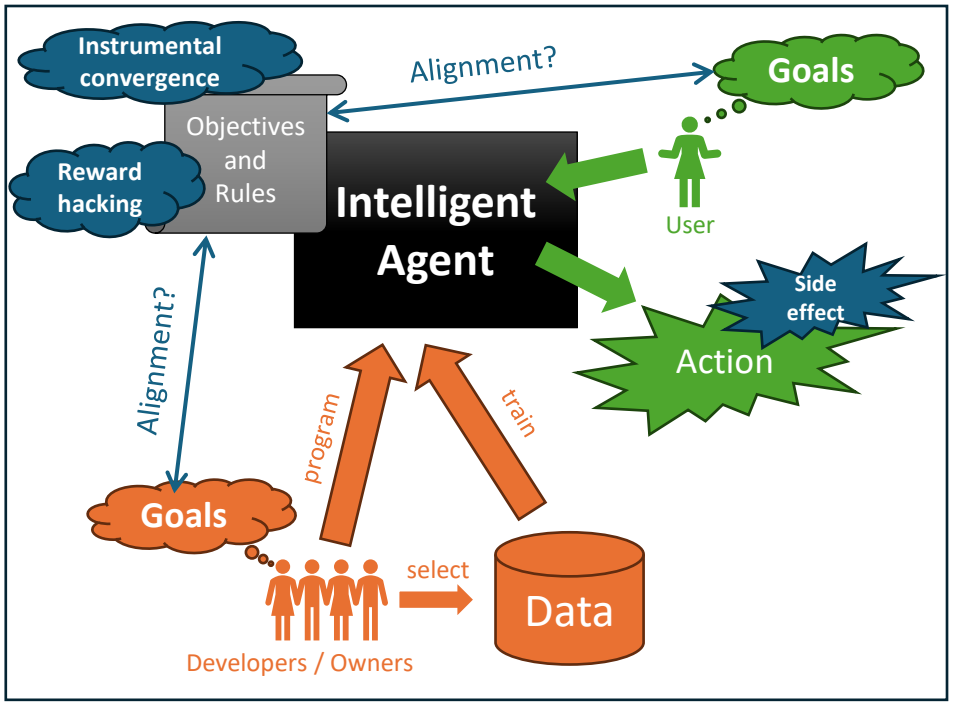


Agent interacting with the environment
[Artificial Intelligence: A Modern Approach, Editions 1-3]

AI Safety and Optimizers

Intelligent Agents are “optimizers!”

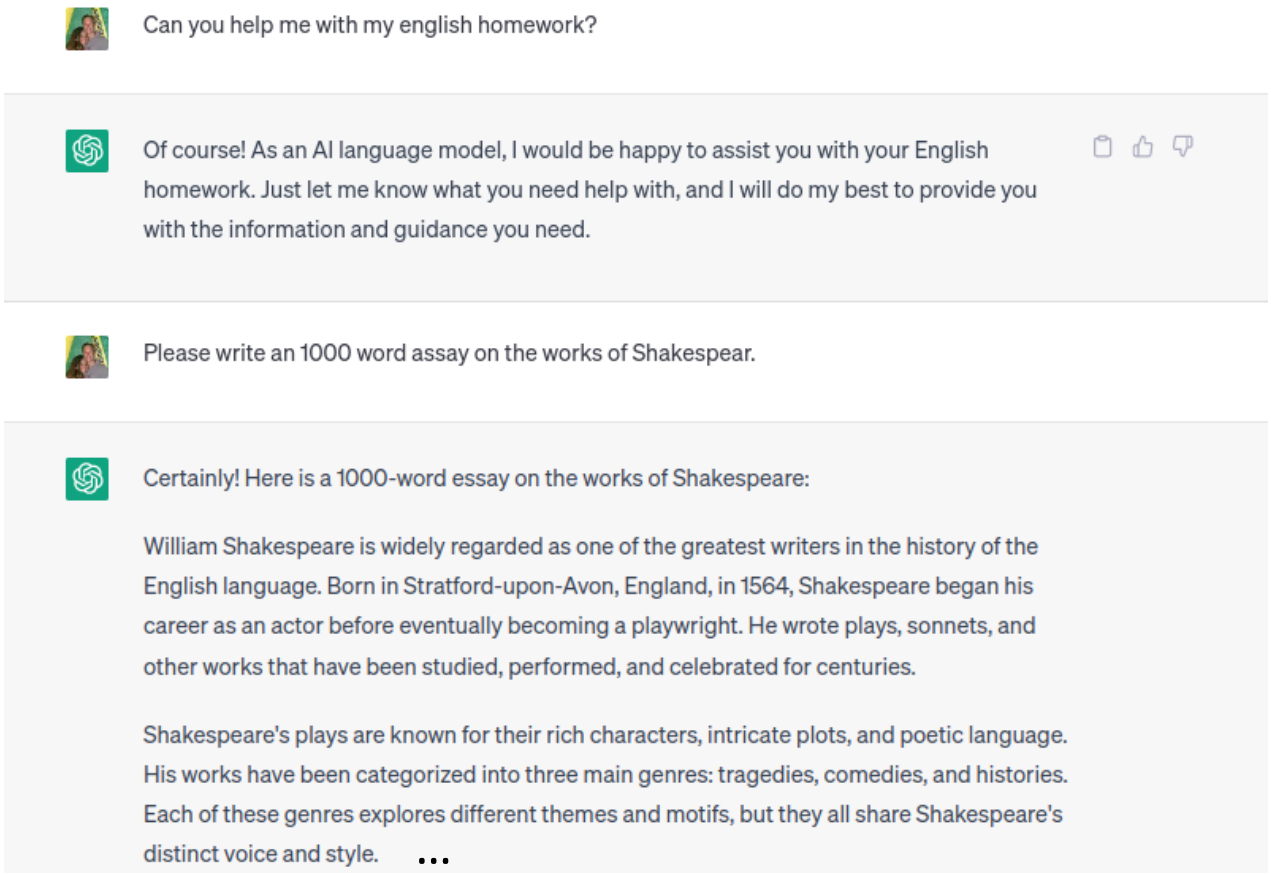
- **Goal/reward alignment**: How do we specify a robust objective function? Whose objectives are used?
- **Reward hacking:** The AI learns to exploit unintended side effects to get a high “score” without solving the objective. AI needs to follow social norms.
- **Instrumental convergence**: All intelligent 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.


Case Study: LLMs

Example Conversation With a Large Language Model (LLM)



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

What are the LLMs

Percepts?

Actions?

Objectives?

Large Language Models (LLMs)

AGI?

"A large language model (LLM) is a computational model notable for its ability to achieve **general-purpose language generation** and other natural language processing tasks such as classification

How?

…

LLMs acquire these abilities **by learning statistical relationships** from vast amounts of text during a computationally intensive self-supervised and semi-supervised training process.

LLMs can be used for text generation, a form of generative AI, by taking an input text and **repeatedly predicting the next token or word**."

Is this intelligent?

Source: Large language model – Wikipedia
https://en.wikipedia.org/wiki/Large_language_model

How do Large Language Models fit into the AI Framework in this Course?

think like a human?

act like a human?

think rationally?

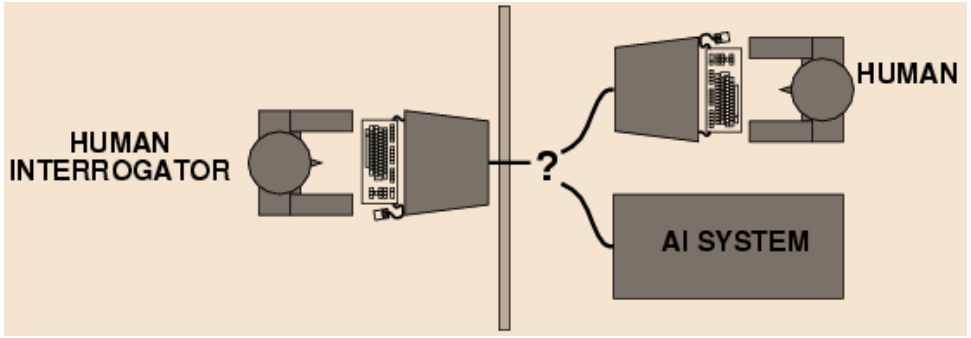
act rationally.

What do LLMs do?
Do LLMs act rationally?

Ask a chatbot if it
- acts rational
- Is an intelligent agent

Turing Test: Large Language Models (LLMs)

act like a human?



Would a modern LLM pass the Turing Test?

- Would you be fooled?
- Why does it or does it not pass your test?
- What does this mean for artificial general intelligence (AGI) or narrow AI?

How do we currently test the performance of LLMs?

- See : Open LLM Leaderboard (Hugging Face)

How do LLMs relate to this:

Chinese Room Argument

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

The AI Effect: AI gets no respect?

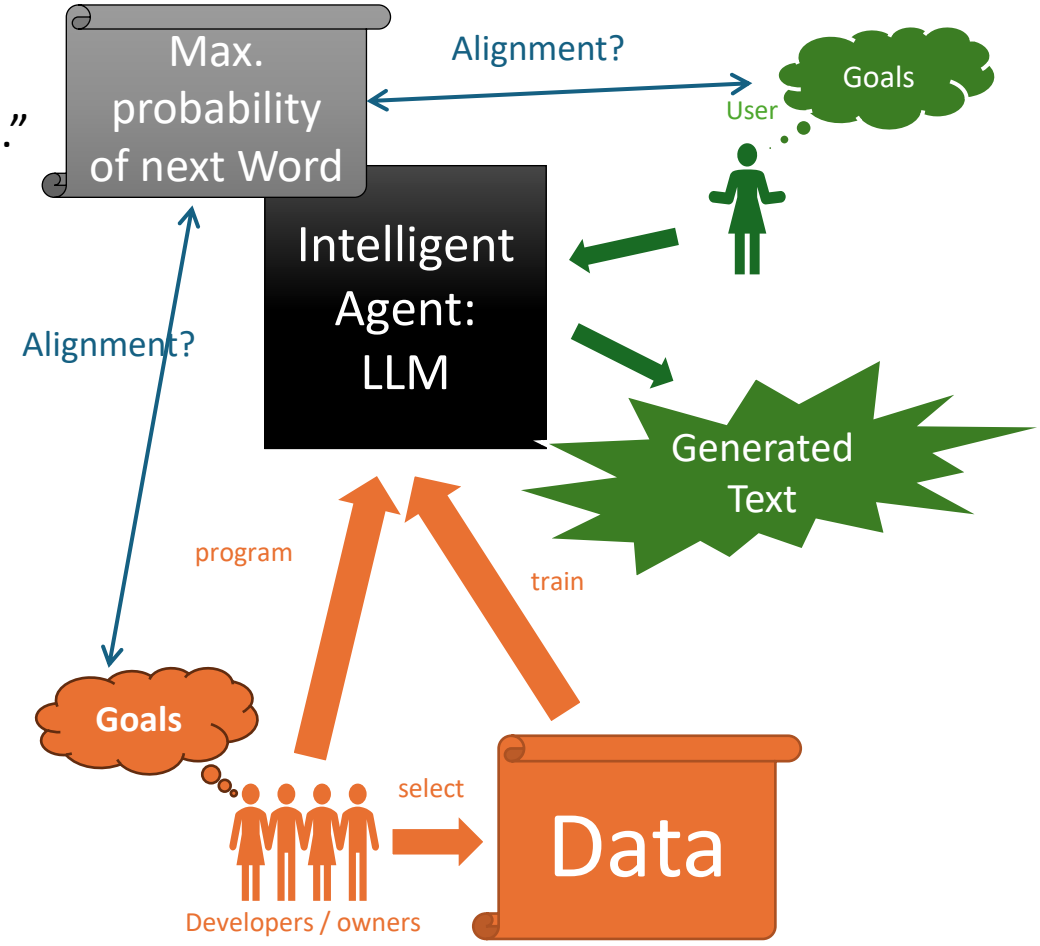
- How do you think LLMs will affect the value of being able to write **assays** as taught in high school?
- LLMs write computer **code**. What does this mean for the value of learning to code?
- When should students be allowed to use the following tools? Give reasons for your decision.
  - A pocket calculator
  - LLMs (to answer homework questions and write assays)
  - LLMs to write or support writing code

AI Safety

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

How are LLMs affected by:

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



Administration Priorities The Record

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.

Some important points:

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

- Should the use of LLMs be regulated?
- How?
- What about copyright?

Conclusion

- LLMs are a powerful new generative AI technology which many applications.

- Unfortunately, there are many open questions. For example:
  - How do LLMs reason and what are the **limits**?
  - How do we make sure that LLMs generate factually **correct output**?
  - How do we fairly **compensate** the people who create the data that is used to train LLMs?
  - How do we use LLMs in **learning**, so human learning is not compromised?