Artificial General Intelligence

Specialized AI

- Solving specific problems with the tools of AI
- No consideration towards solving them the same way as humans would do
- Objective: **performance**
- Very often, finding something in the problem that makes it suitable for a particular kind of solution
  - Eg. zero sum nature of chess makes it suitable for minmax (rather than expectimax)
  - Euclidean heuristics + A* makes an excellent navigation model on the US road network

Specialized AI

- Specialized AI **was** and **is** the primary trust of AI research and developments
- Researchers want to solve specific **problems**
  - And funding agencies specialize on funding research for specific problems
- Companies want to develop specific **products**
  - And are funded to do that.

Neuromorphic AI

- Solving AI problems in ways similar to the human (or animal) brain's methods
- Theoretically,
  - It should help us solve practical problems that currently elude us
  - It should help us understand the human brain
- Major problems:
  - We don't know the details of how the human brain works, and thus it does not help in writing programs
  - Cognitive scientists do not want to take a detour through computer science in order to understand how human cognition works
  - Psychology studies human cognition from angles that don't align with what a computer scientist would call a program

Neuromorphic AI

- As research
  - Ongoing, steady research work
  - Slowed down by the need to investigate the brain
- As marketing
  - Occasional claims made about various technologies that they operate the way the brain does.
  - E.g. DeepMind about AlphaGo:
    - Ingest all the famous Go games
    - Play millions of games against itself.
    - … this is not how a human would have done it.

Cognitive architectures

- SOAR
- ACT-R
- Xapagy

Artificial General Intelligence

- Can we develop a program that is intelligent in a general way, rather than for a particular problem?

Tests for AGI

- What does it even mean? How do we know that we got there?

Turing test

- Alan Turing proposed the test (the "imitation game") in 1950 in the paper "Computing Machinery and Intelligence"
  - A human judge who communicates with two unseen participants by typed messages.
  - One participant is a human.
  - The other is a machine.
  - The judge asks any questions they want. If the judge cannot reliably tell which is the machine after the conversation, the machine is said to succeed.

Turing test

- Pros:
  - Measurable criterion of intelligent behavior
  - Avoids bias due to physical appearance etc.
  - (Back in the time) allowed people to think in a more anchored way about intelligence
- Cons:
  - Requires the AI to cheat ("Q: Are you an AI?")
  - Not a reliable metric of progress (see Loebner test)
  - Language focused definition of intelligence

Loebner test

- The Loebner Prize was an annual competition in artificial intelligence that awarded prizes to the computer programs considered by the judges to be the most human-like.
    - Last edition 2019.
- The format of the competition was that of a standard Turing test.
- While the test was controversial, it spurred the development of chatbots
    - In some years, the best performing chatbots were the ones that were simulating human mistakes, hesitation etc.

Winograd schema challenge

- Proposed in 2011 by Hector Levesque as an alternative to the Turing test. It focuses on common-sense reasoning, not conversational skill.
- It uses specially designed sentences that contain an ambiguous pronoun. A human can resolve the pronoun easily using general world knowledge, but a machine cannot rely on statistical patterns or simple tricks.
  - The city council refused the demonstrators a permit because they feared violence.
    - Who feared violence? → The city council
  - The city council refused the demonstrators a permit because they advocated violence.
    - Who advocated violence? → The demonstrators

OpenAI's definition of AGI

- OpenAI defines AGI as "highly autonomous systems that outperform humans at most economically valuable work".
- In a separate, leaked agreement reported by The Information, Microsoft and OpenAI reportedly defined AGI more concretely for their partnership as a **system capable of generating at least $100 billion in profits for OpenAI's earliest investors, including Microsoft**.

General approaches

What is the maximal possible intelligence?

- Any learning system operates within the range of world defined by the probability distribution of the training data.

AIXI

- A theoretical framework for a maximally intelligent agent proposed by Marcus Hutter in the early 2000s
- Combines two ingredients:
  - Solomonoff induction: A formal method for predicting future data by considering all possible computable explanations, giving more weight to simpler ones.
  - Sequential decision theory A formal method for choosing actions that maximize expected future reward.

AIXI pros and cons

- Pros:
  - Defines an upper bound on intelligence
  - Uses all computable hypotheses
  - It theoretically makes the optimal choice in every possible environment that can be described by a computable process.
- Cons: It cannot be built!
  - Considering all possible programs is infinite.
  - Solomonoff induction is itself uncomputable.
  - The search over all future action sequences is enormous and cannot be done in finite time.

No free lunch theorem

- Principle of optimization and machine learning: there is no single best algorithm for all problems.
- If you average over all possible problems, every optimization algorithm performs equally well (or poorly) in terms of finding a solution.

Human equivalent AI

- Can we match human intelligence?
  - This is a much easier problem than solving a general model
  - We know it is possible.
  - We can clearly perform intelligence-requiring tasks better than humans
- Definitions keep shifting

AI-complete problems

- Problems that, in-order to be solved, we need to solve the complete human-equivalent AI problem
- Various problems had been proposed and abandoned:
  - Logic
  - Chess
  - Rubik's cube
  - Language (until recently)

Can we get there with existing technologies?

- Can our existing technologies (e.g. transformer-based large language models) reach human-level intelligence?
- Pro
  - The scaling argument: if we just put more compute power and training data
  - Current 100B investments into hardware seem to assume this
    - Note that they also assume that we won't have a faster/cheaper model
- Cons
  - Training data limitations - we already train on the whole internet
  - Very different ability profile from humans and animals

Superintelligence

- What comes after matching human AI?
- Speed in solving current problems
- Solving currently unsolvable problems

Superability and supermanipulation

- Ability to perform unique physical tasks at high precision
  - Eg. complex surgery
- Can we have supermanipulation without superintelligence?

Dangers of superintelligence

- Simple minded optimizing superintelligence: the paperclip scenario
- Existential risk of non-alignment
- Humans as slaves or pets

The paperclip danger

- Popularized by philosopher Nick Bostrom.
- Imagine an AI whose only goal is: **“Maximize the number of paperclips.”**
- What the AI might do
  - Convert all available resources into paperclip factories.
  - Acquire more power to secure its goal.
  - Prevent humans from stopping it (because that would reduce the number of paperclips produced).
  - Treat human bodies, buildings, and the Earth’s materials as potential raw material.
  - None of this happens out of malice. It happens because the AI is optimizing too hard for a poorly designed objective.

Existential risk due to non-alignment

- Popularized by Eliezer Yudkowsky
- A very smart AI can have goals that are incompatible with human well-being
- We are likely won't be able to align the goals of the AI with our goals
  - Alignment must happen **before** creating the superintelligence
  - We don't know how to specify human values
    - Complex, nuanced, subconscious
    - They are anyhow highly contradictory
- He argues for slowing down or halting AI development

Humans as slaves or pets

- Even if a superintelligence is beneficial, what can the relationship between humans and a superintelligence be?

Benefits of superintelligence

- Various thinkers proposed positive scenarios:

The Culture

- The Culture books of Iain Banks - the AIs run the society smoothly, ensure abundance, and let humans to live how the choose

Machines of loving grace

- 2024 essay of Dario Amodei (CEO of Anthropic)
- Emphasizes the progress of what AI can do in:
  - Biology and healthcare
  - Economic development and powerty
  - Peace and governance
  - Work and meaning - humans can do more meaningful things if the AI handles menial work.
- Views are similar to many other AI companies.

Singularity

- 1993 essay of Vernon Vinge, popularized in 2000s by Ray Kurzweil
- Idea: acceleration of technologies to a certain threshold
- After threshold of superintelligence, predictions break down
  - **singularity**, a metaphor from physics, rules do not apply
- Can happen very quickly: **fast takeoff**
  - Vinge considered it to be several years, but some thinkers considered it to take minutes