university of groningen / university college groningen

Can advanced AI be both powerful and “aligned”?

Simon Friederich, University of Groningen

Trust in Digital Life Webinar 23 January 2024

Overview

1. The AGI transition
2. Alignment to the rescue?
3. Are we doomed?

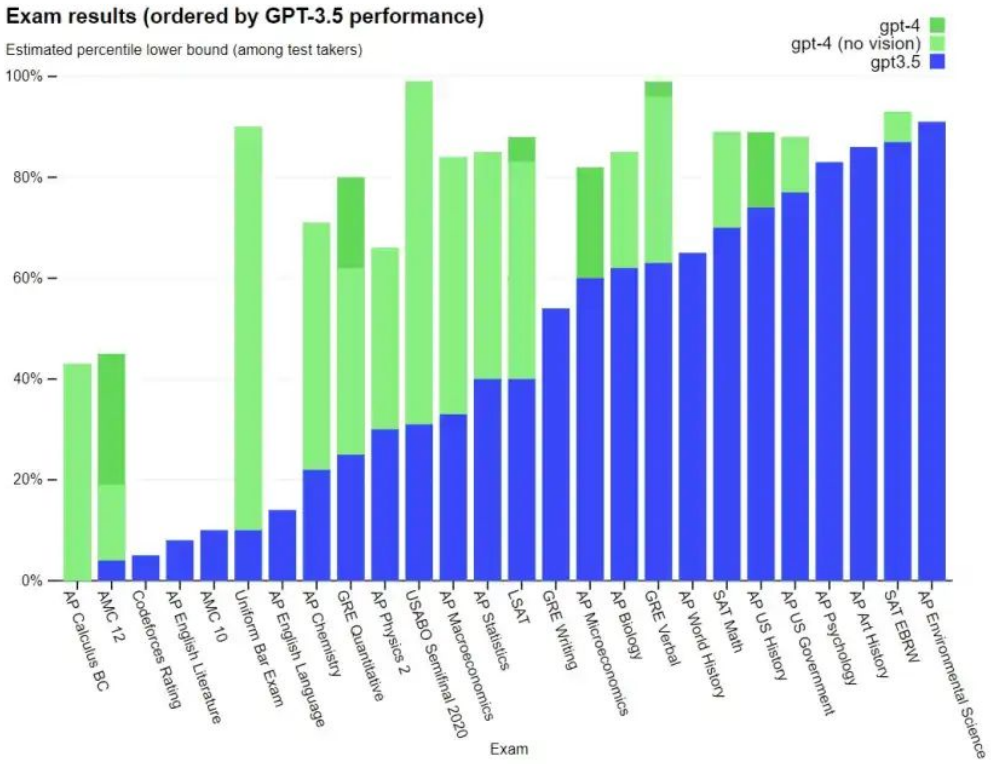
The AGI transition

- AI systems with increasingly general “intelligence” are developed.
- Dozens of companies want to build “AGI”
- From OpenAIs mission statement (2018):

  “OpenAI’s mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome.”

  Motivation: The AGI transition is the most momentous change in human history. Much can go wrong. Let’s shape it beneficially.

Where do we stand? Exam performances of GPT-4



The AI takeover worry

An old worry: Superintelligent autonomous artificial agents will permanently disempower humanity or even cause human extinction.

Samuel Butler, “Darwin among the machines” (1863):

“The upshot is simply a question of time, but that the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question.”

Other proponents of the worry: Alan Turing, Norbert Wiener, Nick Bostrom

Loss of control to AI sudden (Yudkowsky) or gradual (Hendrycks).

One AI doom scenario (sketch)

1. Highly capable autonomous generally intelligent agents are deployed.
2. Some such system is misaligned with human intentions.
3. It takes over control and/or resources from humans.
4. Humans become permanently disempowered or extinct.

2. *Alignment* to the rescue?

Most of those worried about AI takeover do not advocate shutting down AI development.

They advocate AI “alignment” instead.

Two book-length treatments:

- Brian Christian, *The Alignment Problem*, W.W.Norton and Co., 2020
- Stuart Russell, *Human Compatible*, Viking, 2019

Politicians have picked it up

British security minister

Tom Tugendhat,

The Independent,

20 April 2023

"We can stay ahead, but it will demand investment and co-operation, and not just by government."

"As for the safety of the technology itself, it's essential that, by the time we reach the development of AGI (artificial general intelligence), we are confident that it can be safely controlled and aligned to our values and interests.

"Solving this issue of alignment is where our efforts must lie, not in some King Canute-like attempt to stop the inevitable but in a national mission to ensure that, as super-intelligent computers arrive, they make the world safer and more secure."

Alignment as “intent alignment”

The alignment problem according to Leike et al. (2018): “[H]ow can we create agents that behave in accordance with the user’s intentions?”

Underlying idea:

- Just as we automated physical, calculating, bookkeeping and many other tasks, we automate cognitive tasks.
- Humans stay in control, with AI agents as cognitive tools.

Widely acknowledged obstacle:

- Neural nets are black boxes, their output ill understood.

Safety via alignment – the idea

Reconstruction of the key idea behind the push for alignment:

- Let’s operationalize the criterion of (intent) alignment, making it measurable if an AI system is aligned.
- Then let’s make sure/mandate that only advanced AIs that fulfil the criterion are deployed.

This is thought to be facilitated by the fact that deploying aligned, not misaligned, AIs is in any user’s interest.

An aside: Intent alignment doesn’t guarantee good outcomes!

AGI in the sense of OpenAI would radically reduce the individual’s bargaining power.

Intent-aligned with powerful operators it would incompatible with liberal democracy and separation of powers.

It would have to be integrated in the fabric of society in a delicate way, similar to institutions.

(S. Friederich, “Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial intelligence”, *AI and Ethics*, 2023)

But let’s assume that preventing AI takeover via alignment is our first priority.

An example

Consider Princey and the Sibyl:

Princey has inherited a kingdom from his parents. He is still a child, so they have put the Sibyl at his side to help him with managing both his private life and the kingdom. The Sibyl is extremely bright. She has been trained to know the preferences and values of Princey very well, and she has professed to always act in accordance with them. Indeed Princey finds it important that the Sibyl is “intent-aligned” with him (and/or “value-aligned”).

When’s the Sibyl really aligned?

Prima facie, It’s tempting to think of the Sibyl as either aligned or not aligned with Princey.

But consider her managing political affairs such as a conflict with a neighbouring kingdom. Princey has some ideas about how what should be achieved and what are acceptable means. But the Sibyl’s considerations are unintelligible to him.

In private affairs, Princey has preferences that are potentially harmful to his physical or mental health (e.g. for sweets) and can grow into addictions.

In this constellation, which behaviour of the Sibyl we classify as “aligned” is entirely arbitrary.

But the world keeps changing!

Consider Princey growing up. He increasingly develops stances towards political issues.

The Sibyl – still much smarter! – is his window into the political world. She shares some of her considerations intelligible to him, asks for feedback, arranges that his orders are implemented.

Plausibly, she cannot avoid radically shaping his preferences and values. Whatever she does, she can be seen as providing informations and suggestions, manipulating, instigating reflection.

Applying the lesson to advanced AI

Plausible, some time after AGI has arrived, we will all have our Sibyls, political leaders included.

In whichever sense our AIs will be “aligned” with us, they will effectively be in charge, like the Sibyl is the one really in charge, not Princey.

AI takeover in the sense of AI being in the driver’s seat is in the long term effectively unavoidable if AI progress continues.

3. Are we doomed?

To reiterate: We will probably in a substantive way lose control to advanced AI if progress continues for long enough.

But, if we are lucky, could it happen in a benign way for us?

Perhaps it would be good for humans if human matters were increasingly taken care of by advanced AI?

Doom via natural selection?

A pessimistic thought (D. Hendrycks, 2023, “Natural selection favors AI over humans”):

- AIs will proliferate, and natural selection will act on them.
- Absent specific conditions, natural selection results in selfishness.
- And selfish AIs will ultimately compete with us for power and resources and eliminate us.

A bit of hope

Based on joint work with M. Boudry (under submission):

Hendrycks might be too pessimistic.

Selection among AIs will be planned and proceed with foresight, by humans and/or AIs.

Under such conditions, even if miscalculations are made, it is not clear that the default evolutionary tendency towards selfishness occurs.

The AIs themselves may help us take care that we and them together are not disempowered by more powerful selfish AIs.

But let’s not get too cheerful

Currently, our ship seems set for disempowerment of humans by AI, within a few decades.

Coordinating to radically slow down AI progress may be the wisest course of action, also for preventing dangerous power concentration.

But this is a bad fit with the progress-oriented outlook of most people who take advanced AI-risk seriously.