...we had better be quite sure that the purpose put into the machine is the purpose which we really desire

not merely a colourful imitation of it

The alignment problem

Musings on the Alignment Problem

Jan Leike

What it is

Today's alignment problems

The hard problem of alignment

Capability vs Alignment

To construct a performant AI system that will achieve some intended task, we need:

Capability	Alignment
The AI system could do the intended task	The AI system does the intended task as well as it can

When an AI system doesn't achieve an intended task, this is because of:

- Failure of capability
- Failure of alignment
- Failure of both capability and alignment

We care about alignment with human intentions

Intended task: what the human wanted the system to do

Reference:
J. Leike, https://aligned.substack.com/p/what-is-alignment (2022)

Familiar Examples

Capability problems

Many problems AI systems have had to date have been capability problems

The systems have simply lacked the capability to perform the intended task

Examples

You cannot jump 10m high

You cannot do arithmetic in 1 nanosecond

Alignment problems

A misaligned system "doesn't play on your team"

It could be playing against you, but it need not be - often it is just playing a different game

Examples

Someone jumps in front of you in a queue

You have to watch an ad before a video

A company sends you unsolicited promotions you don't want

Not capability problems

Person/system chose not to do what you wanted them to do

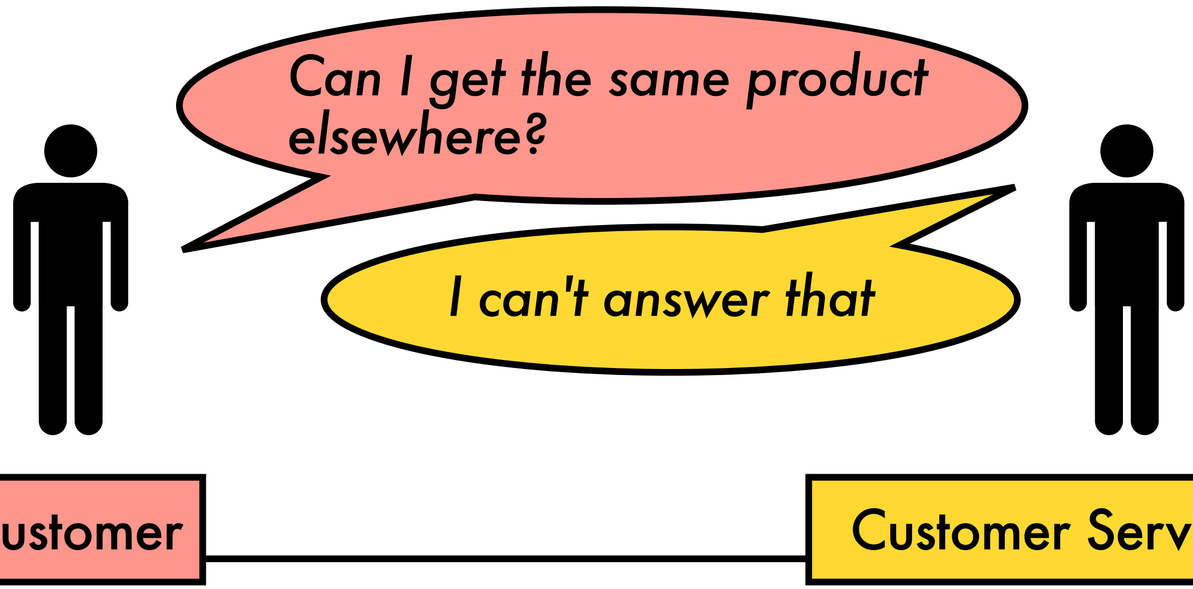
Reference:
J. Leike, https://aligned.substack.com/p/what-is-alignment (2022)
(high jump logo) https://commons.wikimedia.org/wiki/File:High_jump_pictogram.svg
(brain logo) https://commons.wikimedia.org/wiki/File:Noun_brainstorming_1325497.svg

Can We Disentangle Capability And Alignment?

How can we distinguish a capability problem from an alignment problem?

It is very difficult!

To demonstrate misalignment, we need to show that the system could have achieved the task



Capability problem

Doesn't know the answer

Alignment problem

Instructed not to answer by employer

Without detective work, it is hard to tell which it is

Reference:
J. Leike, https://aligned.substack.com/p/what-is-alignment (2022)

Current Alignment Problems

Clearest examples of misalignment problems in today's AI arise in Large Language Models (LLMs)

LLMs fail to act in a way that matches our intentions (either explicit or implicit)

Explicit Intentions	Implicit Intentions
Can be communicated via instructions	Not stated explicitly and difficult to enumerate
Examples	Examples
Translate this passage to French	Don't lie
Summarise this text in two sentences	Don't use toxic language
	Don't give harmful advice

One approach to tackling alignment in current systems: finetune on curated data InstructGPT

If we can't align current AI systems, our alignment methods are seriously flawed

Reference:
J. Leike, https://aligned.substack.com/p/what-is-alignment (2022)
(InstructGPT) L. Ouyang et al., "Training language models to follow instructions with human feedback", arxiv (2022)

The Hard Problem Of Alignment

Current AI systems raise different problems to future AI systems that are smarter than us

For now, humans can still evaluate AI system performance

The hard problem of alignment:

How do we align systems on tasks that are difficult for humans to evaluate?

AI progress will allow the use of AI systems on increasingly difficult tasks

This will make it more difficult for humans to assess whether AI behaviour accords with their intent

We won't be able to use current techniques like Reinforcement Learning from Human Feedback

The hard problem is also where the stakes are highest

There'll be tremendous economic pressure to use AI that can do hard tasks better than humans

If the alignment problem is not solved, these systems will not do the tasks as humans intend

References:
J. Leike, https://aligned.substack.com/p/what-is-alignment (2022)
J. Leike, M. Martic, S. Legg, https://www.deepmind.com/blog/learning-through-human-feedback (2017)

This will have unintended consequences....