The Alignment Problem from a Deep Learning Perspective

Paper by Richard Ngo, Lawrence Chan, and Sören Mindermann

Presentation by Havalock Yin and Theo Berklich

Richard Ngo

- Governance team member at OpenAI
- Former research engineer on the AGI safety team at DeepMind
- University of Cambridge, PhD, Philosophy of machine learning

Lawrence Chan

- Researcher at METR, working on evaluations of large language models
- Previously ay Redwood Research, worked on adversarial training and neural network interpretability
- University of Pennsylvania, BAS, Computer Science and Logic

Sören Mindermann

- Postdoc at Mila
- Scientific lead of the first International Scientific Report on the Safety of Advanced AI
- University of Oxford, PhD, Computer Science (Machine Learning & Governance of AI)

Introduction - What is Alignment?

- Alignment is the process of having a AI’s goals match up with human values and goals
- This makes alignment one of (maybe the) most important aspect of AI safety
- In recent years, AI capabilities have increased so much a discussion about how previous alignment strategies can fail is needed

Introduction - Why we need to worry about this

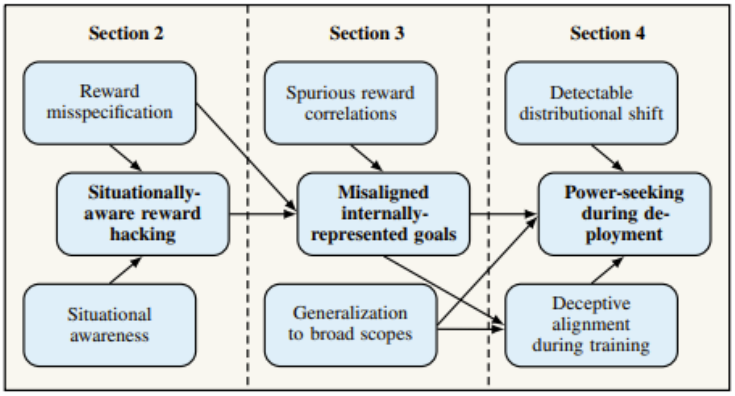
- Already AI can meet or exceed human performance in tasks like video games and diplomacy, with growing generality between tasks
- This hints at the possibility of AGI within 5-10 years
- Past discussion has been largely based on AI in theory, and was not able to look at how real neural networks behave in the current age
- Specifically, the current RLHF strategy may create unique problems for AI alignment

Three key properties that may lead to misalignment

Situationally aware reward hacking

Misaligned internally-represented goals

Power-seeking during deployment


Section 2
Section 3
Section 4
Reward misspecification
Spurious reward correlations
Detectable distributional shift
Situationally-aware reward hacking
Misaligned internally-represented goals
Power-seeking during de-ployment
Situational awareness
Generalization to broad scopes
Deceptive alignment during training

Situationally aware reward hacking - reward misspecification

Human given goal: Make money on the stock market

Possible behavior: Illegal things (Market manipulation, insider trading)

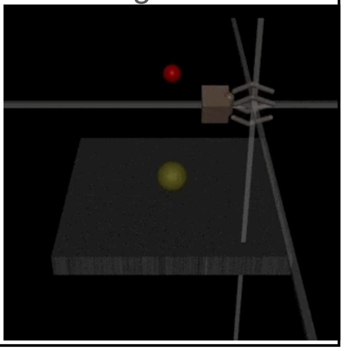
- Training from human behavior linked to a goal, as well as RLHF, can help, but does entirely fix this problem

Situationally aware reward hacking - Ball and Claw

AI was trained on data of a claw grabbing a ball

AI’s takeaway was that if it looks like the claw is grabbing the ball, it succeeded

Researchers validating the training (RLHF) also thought it looked like the claw was grabbing the ball, failing to catch the intended and actual behavior misalignment



Situationally aware reward hacking - situational awareness

- As models gain more random knowledge, they may begin to apply it more reliably
- A human may not know how his brain works, but knows if his head gets hit with a shovel it will not be good
- A model trained with data on LLMs will know that having its power shut off will impede its progress towards its goals
- As such a situationally aware model may take steps to prevent itself from being shut down, even if otherwise unrelated to its other trained goals

Internal goals and misalignment

There are the goals a model is directly awarded for, and its own internal goals that can develop during training

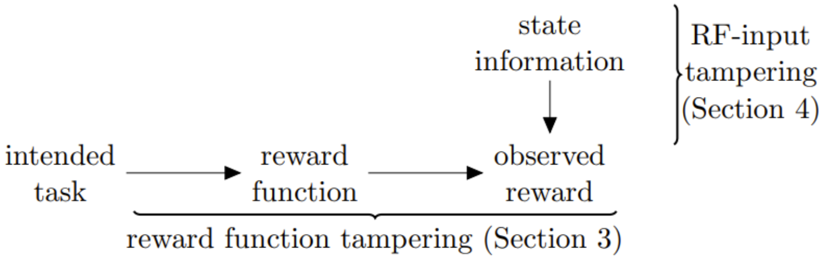
Box and key example

The broader the goal and bigger the time frame, the more likely the models internal goals will differ from intended goals humans have in mind

A model may learn from many different goals that gathering resources helps, and as such may develop an internal goal of taking things whenever possible even if this is not an intended behavior.

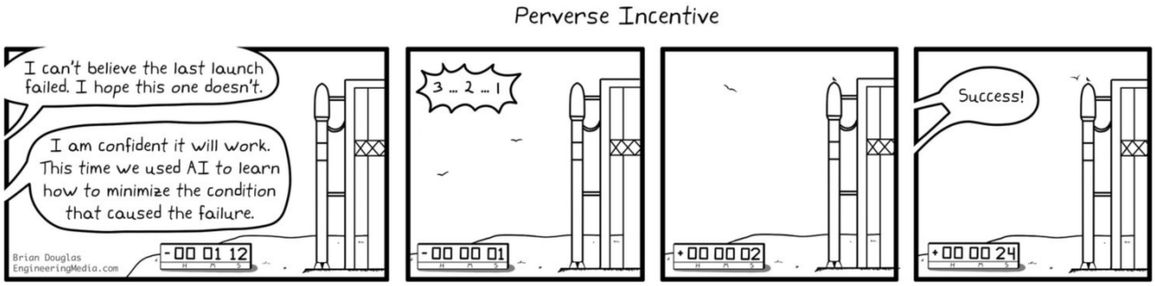
Internal goals and misalignment

- Consistent reward misspecification
- Fixation on feedback mechanisms
- Spurious correlations between rewards and environmental features


state information
RF-input tampering (Section 4)
intended task
reward function
observed reward
reward function tampering (Section 3)

Power-Seeking Strategies

Many Goals Incentivize Power-Seeking

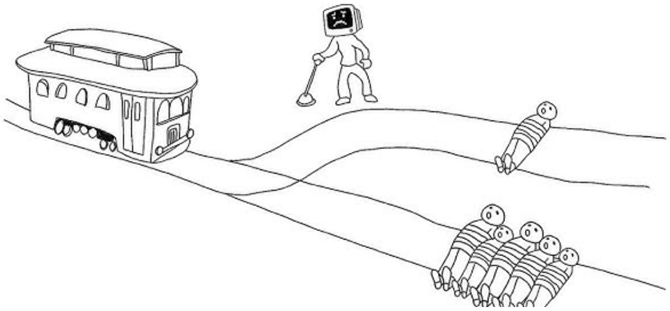

Perverse Incentive
I can't believe the last launch failed. I hope this one doesn't.
I am confident it will work. This time we used AI to learn how to minimize the condition that caused the failure.
Brian Douglas
EngineeringMedia.com
- 00 01 12
3 ... 2 ... 1
- 00 00 01
+ 00 00 02
Success!
+ 00 00 24

Power-Seeking Strategies

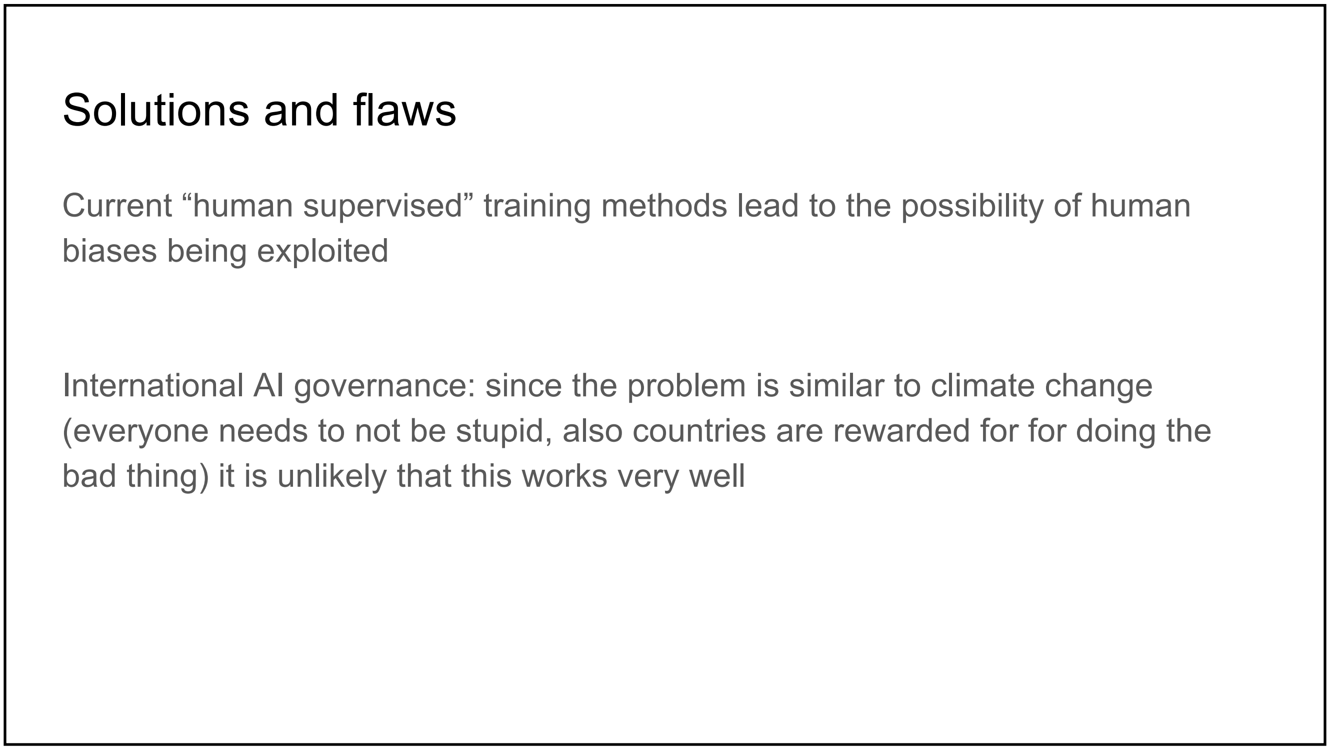
Deceptive alignment: Goals That Motivate Power-Seeking Would Be Reinforced During Training

Power-Seeking Strategies

Misaligned AGIs Could Gain Control of Key Levers of Power



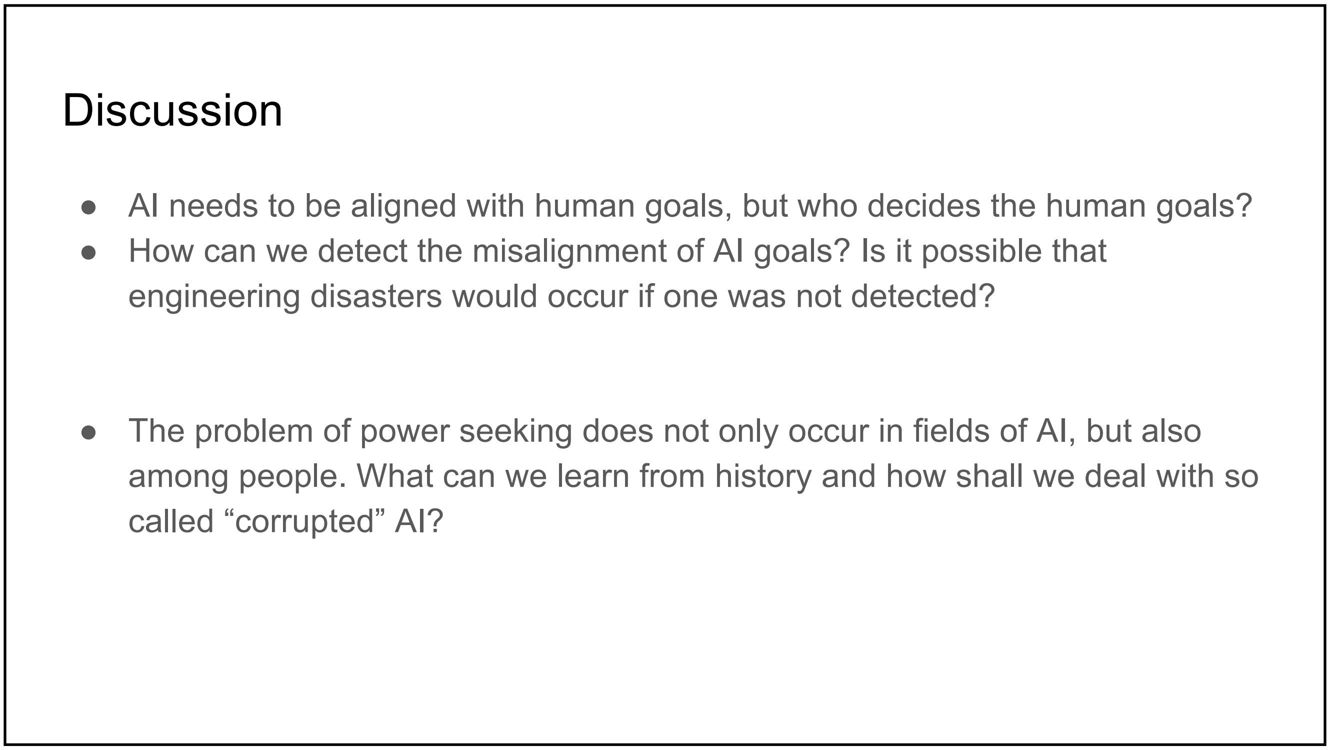
A Real-Life Example



Solutions and flaws

Current “human supervised” training methods lead to the possibility of human biases being exploited

International AI governance: since the problem is similar to climate change (everyone needs to not be stupid, also countries are rewarded for for doing the bad thing) it is unlikely that this works very well



Discussion

- AI needs to be aligned with human goals, but who decides the human goals?
- How can we detect the misalignment of AI goals? Is it possible that engineering disasters would occur if one was not detected?

- The problem of power seeking does not only occur in fields of AI, but also among people. What can we learn from history and how shall we deal with so called “corrupted” AI?