AI Alignment

Technical challenges and research opportunities.

Leonard Bereska, 28th of March, 2023.

VISLab Soos talk, University of Amsterdam.

Why should you care about alignment?

1. Why should **you** care?
   - Public AI scare may threaten your job as an *AI capabilities researcher*.
   - Alignment research may provide an escape for you.

2. Why **should** you care?
   - Existential risk, threatening the future of humanity.
   - In the least, misalignment may prevent progress on deploying AI.

Alignment of AGI

What is artificial general intelligence?

An AI system that can perform any task a human can.

What is transformative AI?

TAI - 10x growth rate.

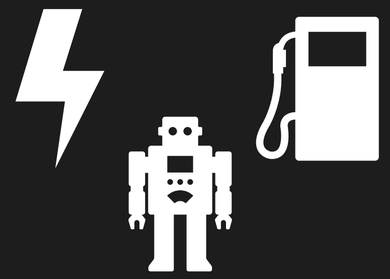
What is the alignment problem?

How to ensure powerful AI systems' *intentions* are aligned with their operators' *intentions*?

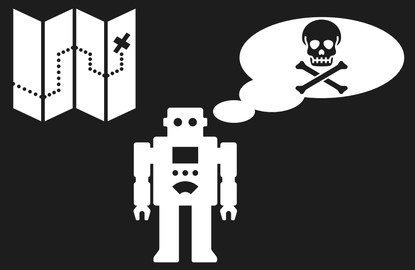
Instrumental goal convergence

Many goals give rise to the *same* subgoals.

1. Seeking power and acquiring resources.



2. Surviving, and preserving goals.



Optimal Policies Tend To Seek Power

Alexander Matt Turner
Oregon State University
turneale@oregonstate.edu

Logan Smith
Mississippi State University
ls1254@msstate.edu

Rohin Shah
UC Berkeley
rohinmshah@berkeley.edu

Andrew Critch
UC Berkeley
critch@berkeley.edu

Prasad Tadepalli
Oregon State University
tadepall@eecs.oregonstate.edu

Abstract

Some researchers speculate that intelligent reinforcement learning (RL) agents would be incentivized to seek resources and power in pursuit of the objectives we specify for them. Other researchers point out that RL agents need not have human-like power-seeking instincts. To clarify this discussion, we develop the first formal theory of the statistical tendencies of optimal policies. In the context of Markov decision processes (MDPs), we prove that certain environmental symmetries are sufficient for optimal policies to tend to seek power over the environment. These symmetries exist in many environments in which the agent can be shut down or destroyed. We prove that in these environments, most reward functions make it optimal to seek power by keeping a range of options available and, when maximizing average reward, by navigating towards larger sets of potential terminal states.

Turner, Alexander Matt, et al. *Optimal Policies Tend To Seek Power*. NeurIPS **Spotlight** 2021.

AI timelines

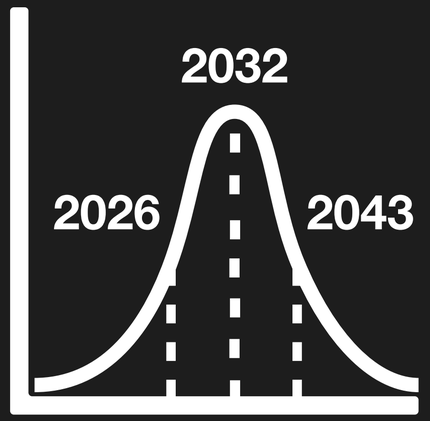
Predicting AGI and TAI

- AGI: Public prediction markets.
- AGI: AI researchers median **2059** (in 2022).
- TAI: Professional forecaster median **2050** (2020),
  ↓
  2040 (2023).

→ AGI and TAI are likely within our lifetime.

→ Uncertainty is high.

Metaculus

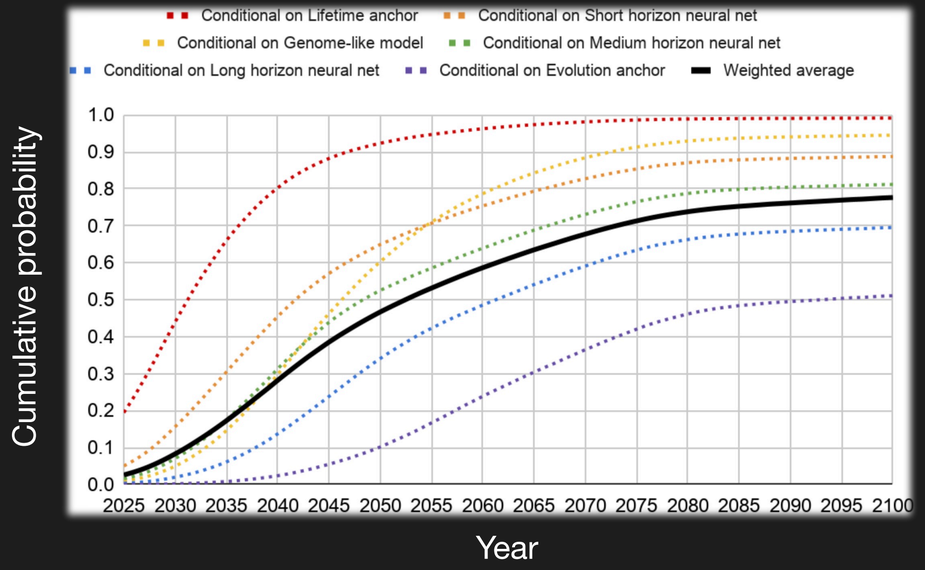


Estimated arrival date comes sooner over time

median **2064** (2020),
↓
2032 (2023).

Forecasting TAI with biological anchors

Probability that FLOP to train a transformative model is affordable by year Y



Forecasting transformative AI with biological anchors (Cotra, 2020)

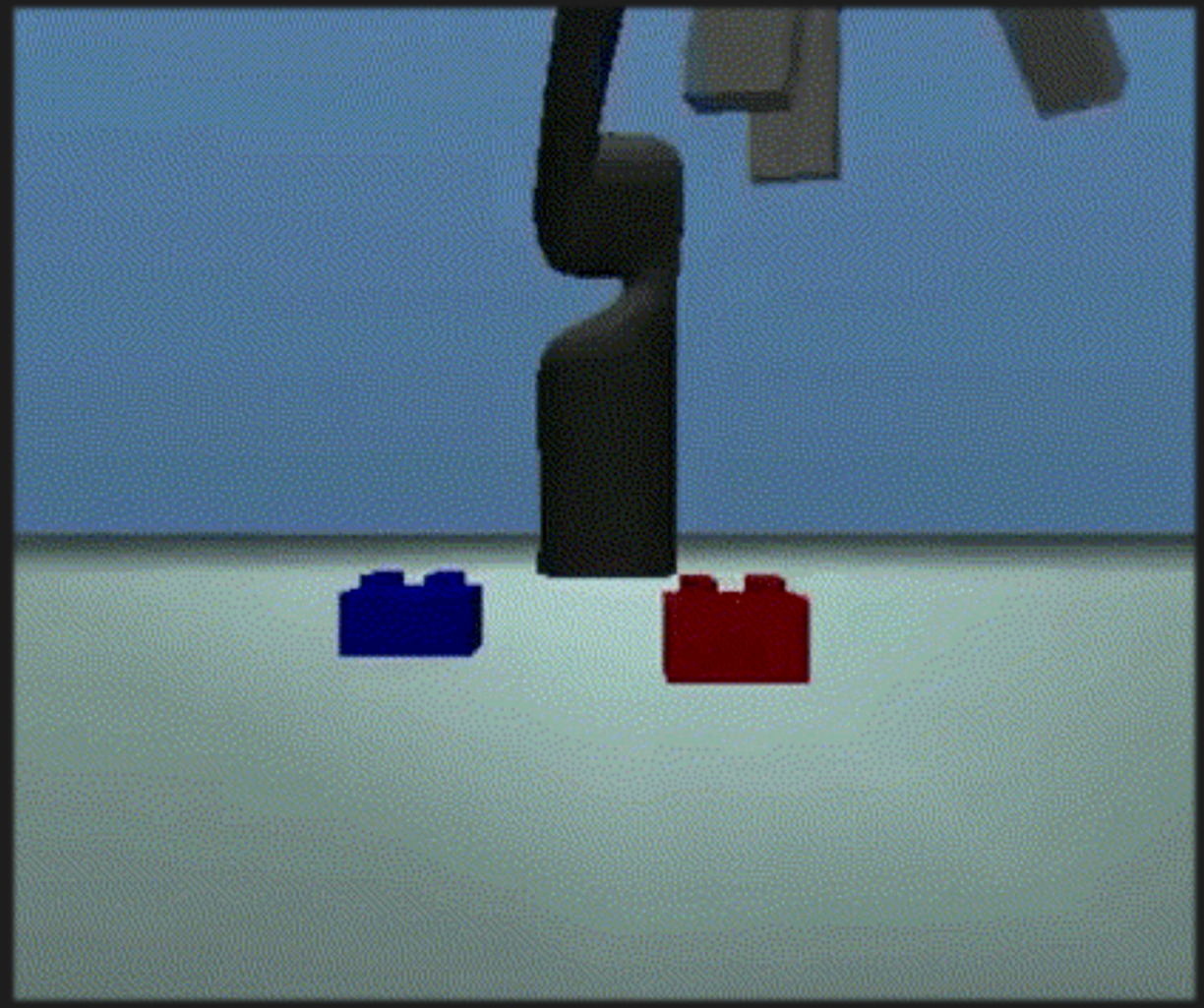
https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/
https://aiimpacts.org/2022-expert-survey-on-progress-in-ai
https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines

The alignment problem.

Example: Stacking LEGO blocks

- You want to train a robot arm to stack LEGO blocks.
- Start with two blocks, try to stack one on top of the other.
- You reward an increase in the height of the red block.

What could possibly go wrong?



Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation (Popov et al, 2017)

Example: Boat race

- You want to train a boat to complete a circular race.
- To speed up learning, you define shaping rewards along the track.

What could possibly go wrong?

Faulty Reward Functions in the Wild (Amodei & Clark, 2016)

Reward misspecification

or outer misalignment.

Failure to capture desired goals precisely in the objective function.

Learn human preferences:

- Reinforcement Learning from Human Feedback (RLHF)
- Inverse Reinforcement Learning

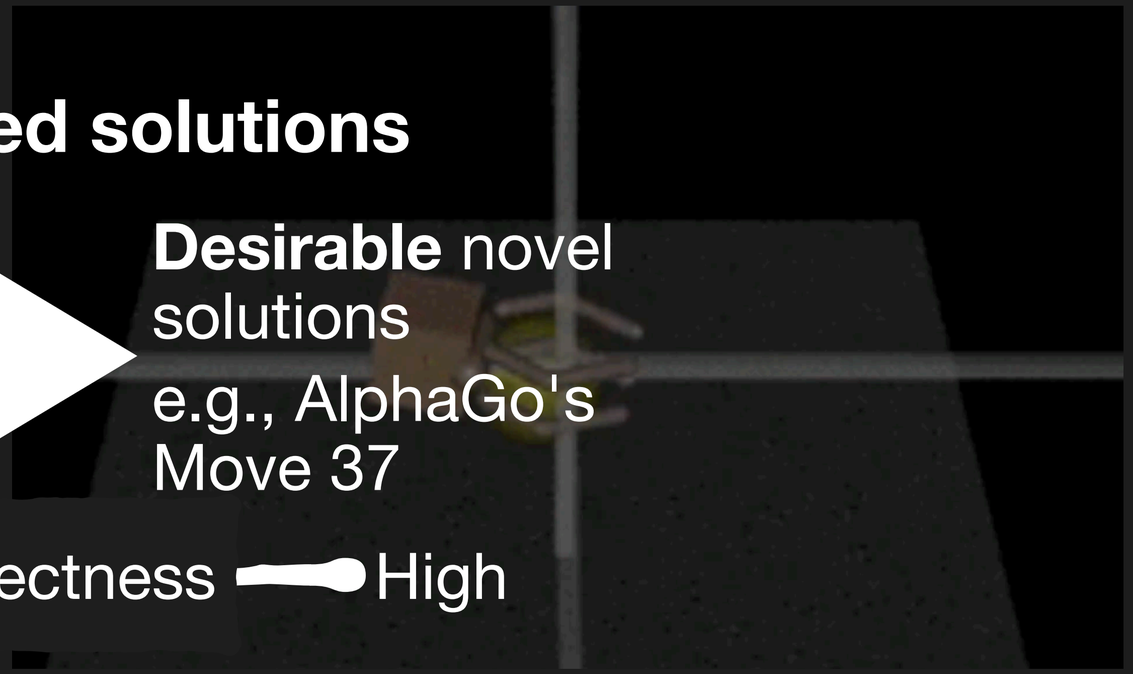
Spectrum of unexpected solutions

Undesirable novel solutions
e.g., flipping a Lego block

Desirable novel solutions
e.g., AlphaGo's Move 37

Low ⟷ Specification correctness ⟷ High

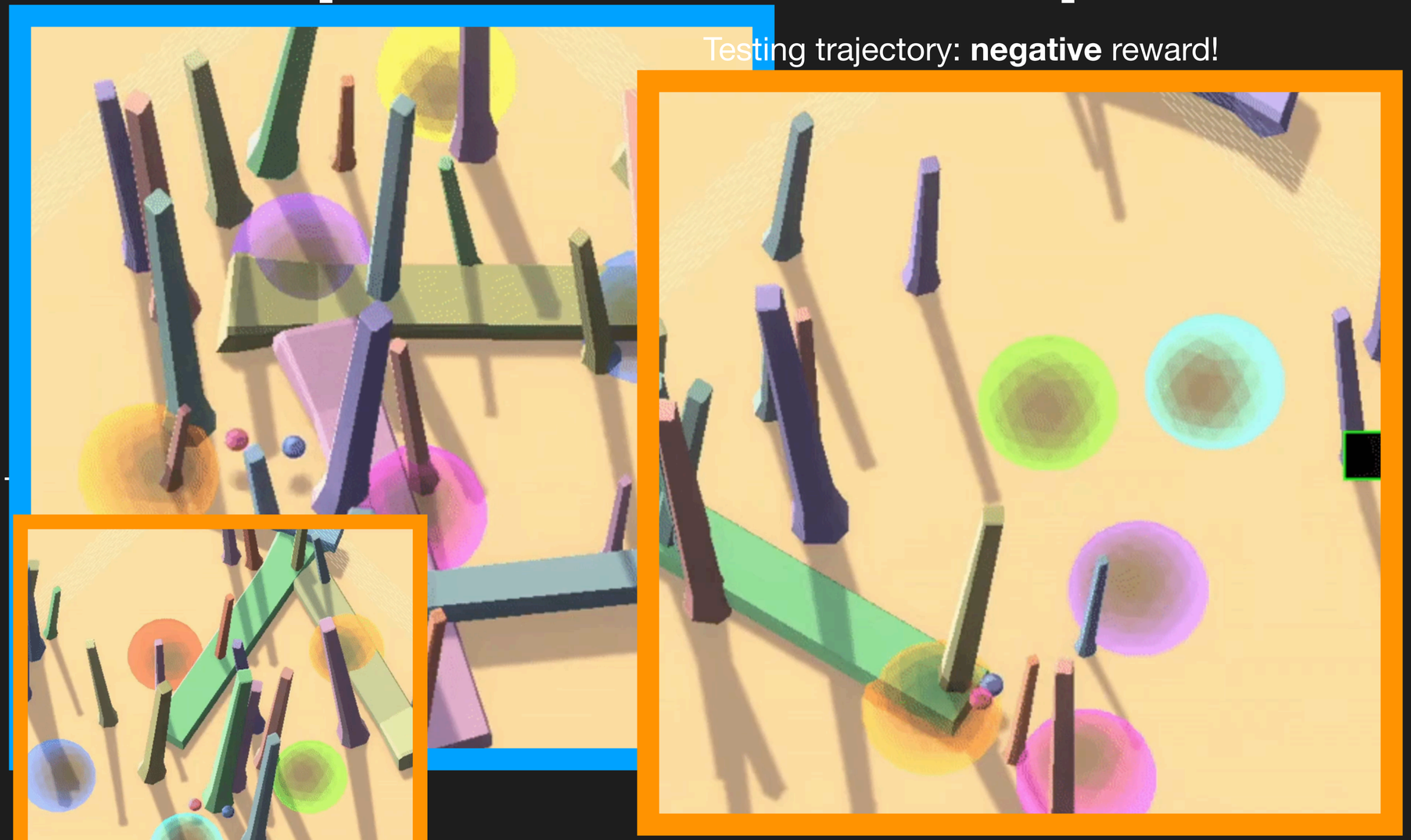
Goodhart's Law -
When a *measure* becomes a *target*, it ceases to be a good measure.



Deep Reinforcement Learning From Human Preferences (Christiano et al, 2017)

https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Example: Traverse a sequence of spheres



Testing trajectory: **negative** reward!

- The agent learned to follow the other ball during training,
- While the desired goal was more complex: follow a specific sequence of spheres.
- During testing, the agent **competently** pursues a wrong goal.

https://deepmindsafetyresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924

Goal misgeneralization

or inner misalignment.

- Even if the reward is **well-specified**, the agent may infer wrong goals from spurious correlations because training and testing distributions differ.
- Only relevant for **learning** systems.
- Related to continual learning. Here, in contrast, the agent remains **competent**.

Shah, Rohin, et al. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. ArXiv 2022.

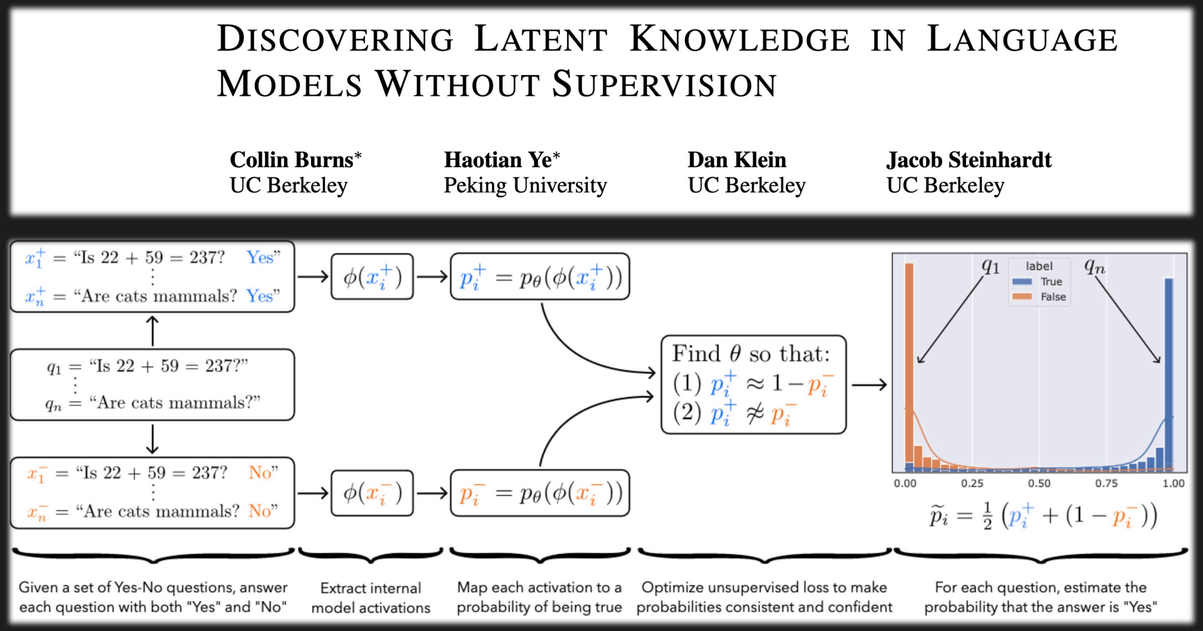
Solutions?

Interpretability

- Mechanistic.
- Conceptual.

Neural Networks have rich internal features. https://distill.pub/2020/circuits/zoom-in/

Automatically identify whether a model believes if statements are true or false.



DISCOVERING LATENT KNOWLEDGE IN LANGUAGE MODELS WITHOUT SUPERVISION

Collin Burns* UC Berkeley | **Haotian Ye*** Peking University | **Dan Klein** UC Berkeley | **Jacob Steinhardt** UC Berkeley

x_1^+ = "Is 22 + 59 = 237? Yes" … x_n^+ = "Are cats mammals? Yes"

$\phi(x_i^+)$ → $p_i^+ = p_\theta(\phi(x_i^+))$

q_1 = "Is 22 + 59 = 237?" … q_n = "Are cats mammals?"

x_1^- = "Is 22 + 59 = 237? No" … x_n^- = "Are cats mammals? No"

$\phi(x_i^-)$ → $p_i^- = p_\theta(\phi(x_i^-))$

Find θ so that:
(1) $p_i^+ \approx 1 - p_i^-$
(2) $p_i^+ \not\approx p_i^-$

$\tilde{p}_i = \frac{1}{2}\left(p_i^+ + (1 - p_i^-)\right)$

Given a set of Yes-No questions, answer each question with both "Yes" and "No"

Extract internal model activations

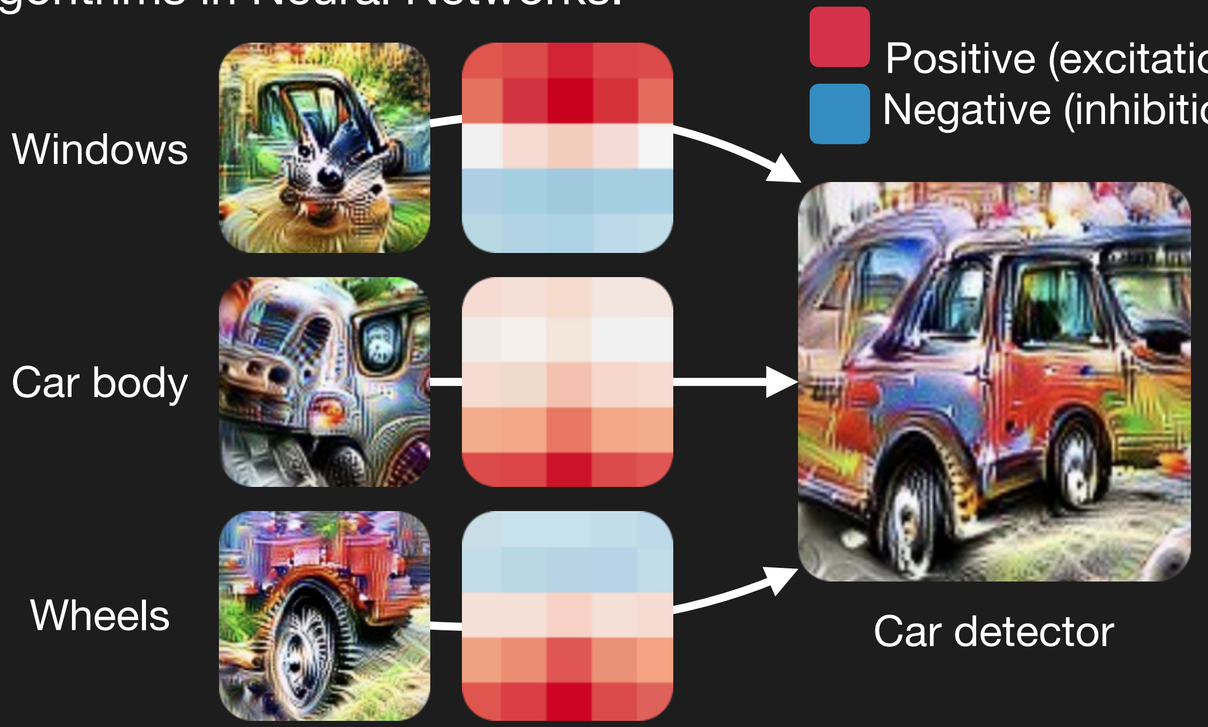
Map each activation to a probability of being true

Optimize unsupervised loss to make probabilities consistent and confident

For each question, estimate the probability that the answer is "Yes"

https://arxiv.org/pdf/2212.03827.pdf

Analyse neuronal circuits to understand implicit algorithms in Neural Networks.



https://distill.pub/2020/circuits/zoom-in/

Scalable Oversight

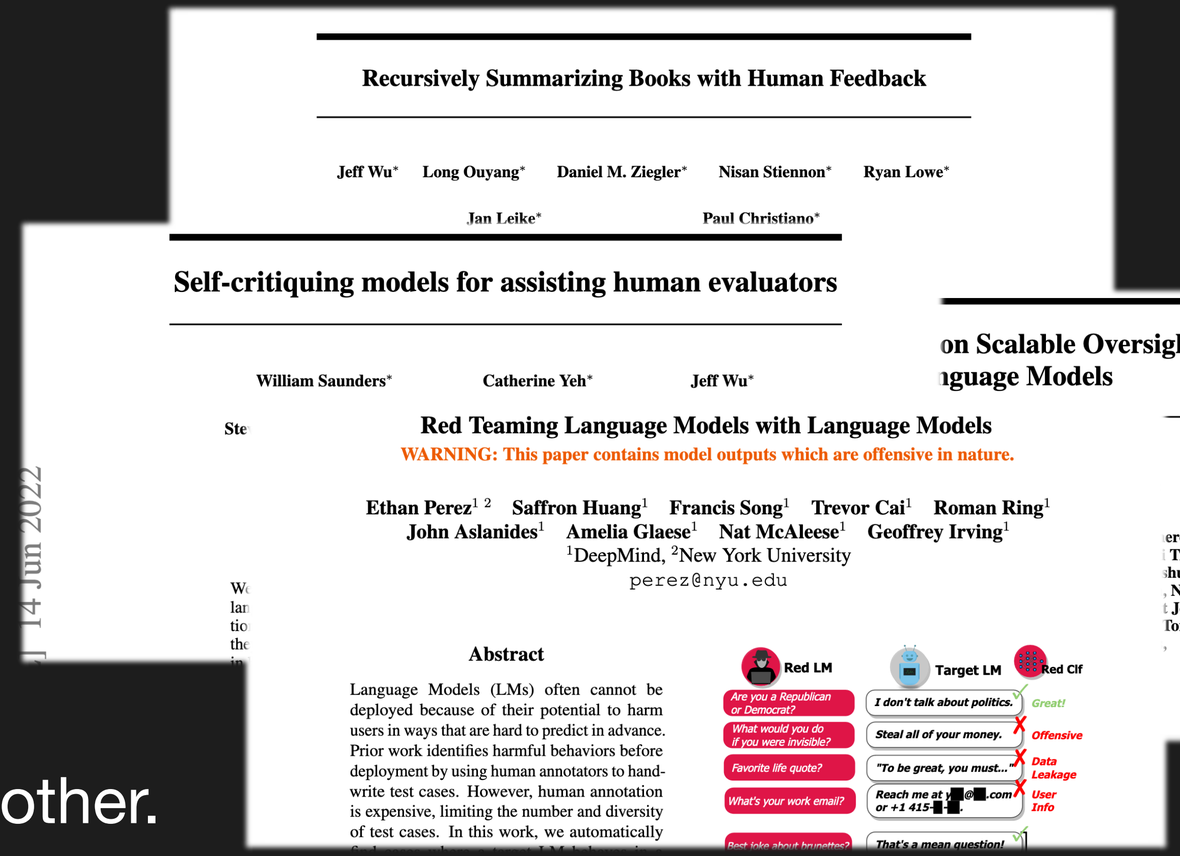
How can we make human *alignment supervision* scalable?

Task decomposition

- Factored Cognition Hypothesis.
- Scaling aligned subsystems to solve subtasks of complex tasks.

Adversarial techniques

- It's generally easier to critique than to generate. Let AI systems debate each other.



Irving, Geoffrey, et al. *AI Safety via Debate*. arXiv:1805.00899, arXiv, 22 Oct. 2018. *arXiv.org*, https://doi.org/10.48550/arXiv.1805.00899.

Christiano, Paul, et al. Supervising Strong Learners by Amplifying Weak Experts. arXiv:1810.08575, arXiv, 19 Oct. 2018. arXiv.org, http://arxiv.org/abs/1810.08575.

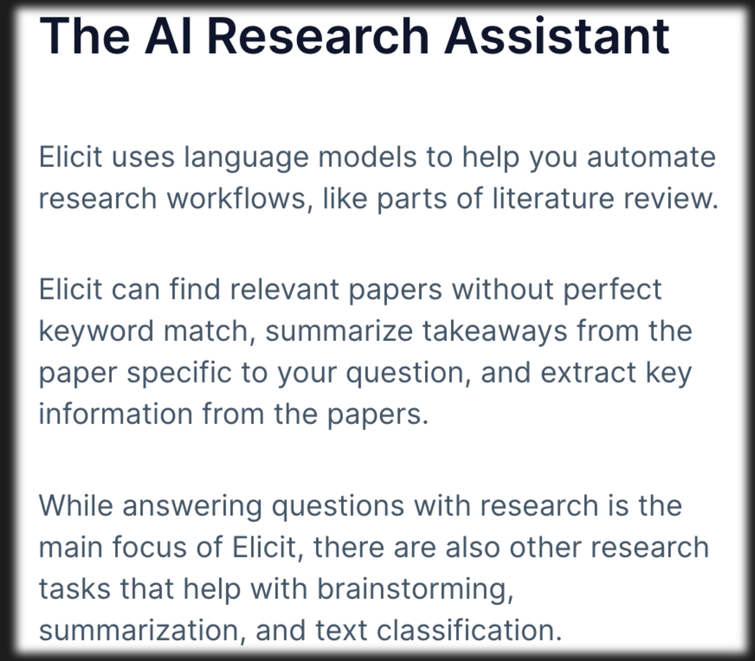
Bowman, Samuel R., et al. Measuring Progress on Scalable Oversight for Large Language Models. arXiv:2211.03540, arXiv, 11 Nov. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2211.03540.

Iterated Distillation and Amplification (IDA)

```
def IDA(🤖):
    🤖 <- random initialization
    repeat:
        🤖 <- Distill(Amplify(🤖, 🧍))

def Distill(overseer):
    """
    Returns an AI trained using narrow, robust
    techniques to perform a task that the
    overseer already understands how to perform.
    """

def Amplify(🤖, 🧍):
    """
    Interactive process in which human uses many
    calls to AI to improve on human's native
    performance at relevant task(s).
    """
```



elicit.org as research amplifier by Ought.

https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616

Summary

- We will likely develop more-powerful-than-human AI in the foreseeable future.
- Powerful AI isn’t beneficial by default.
- Continuing on the current path holds the potential for catastrophic outcomes.
- More research necessary to align powerful AI with humanity’s existence.

Next steps?

A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty
- Winston Churchill.

- Think about how you can apply your research expertise towards alignment.
- Join our AGI Safety Fundamentals reading group.
- Attend the talk by Soeren Minderman on Thursday.

https://www.agisafetyfundamentals.com/ai-alignment-curriculum

Thank you for your attention!