AI Safety Initiative Amsterdam

ELLIS BeNeLux

Leonard Bereska November 7, 2023.

Supervised by Prof. Efstratios Gavves at UvA.

Hey, I’m Leonard!

- Third Year *PhD* Student.
- University of *Amsterdam*.
- Background in *Physics*.
- Working in *AI Safety*.
- Research focus:
  Mechanistic Interpretability of Transformer Models.

Overview

- Why care?
- The Alignment Problem.
- What about LLMs?
- Transparency
- AI Safety Initiatives

Why care?

Why care about AI 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?

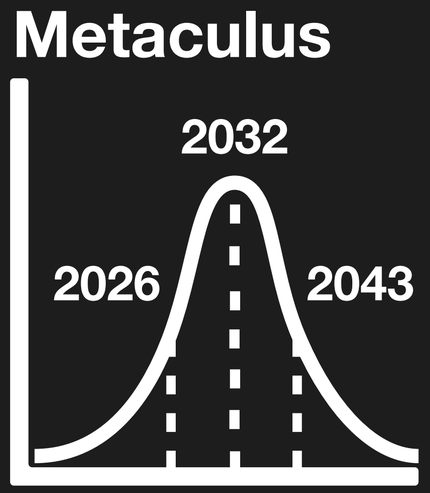
TAI - 10x growth rate.

What is the alignment problem?

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

AI timelines

Predicting AGI and TAI



- AGI: Public prediction markets.
- AGI: AI researchers median **2059** (in 2022).
- TAI: Ajeya Cotra (professional forecaster)

  median **2050** (2020)
  ↓
  2040 (2023).

→ **AGI and TAI are likely within our lifetime.**

→ **Uncertainty is high.**

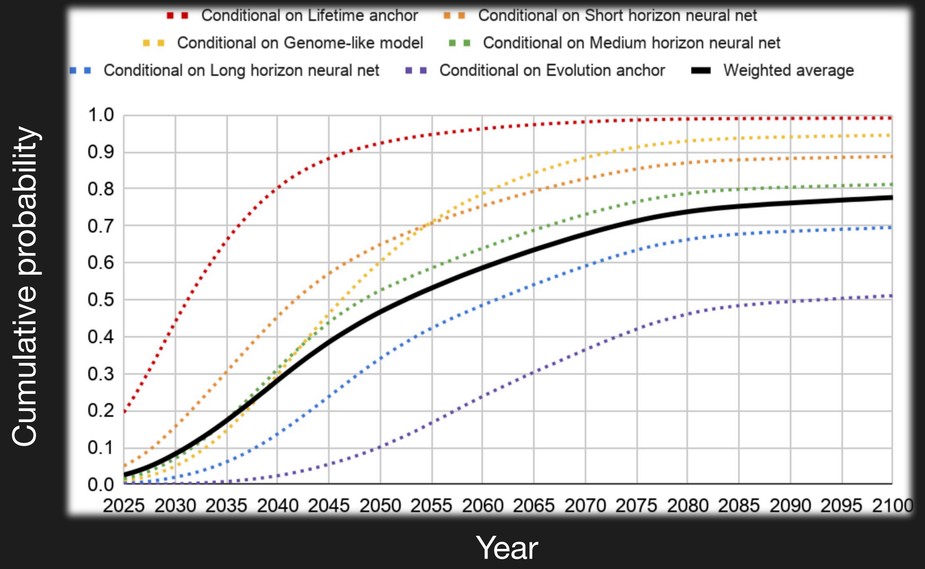
https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/
https://aiimpacts.org/2022-expert-survey-on-progress-in-ai
https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines

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



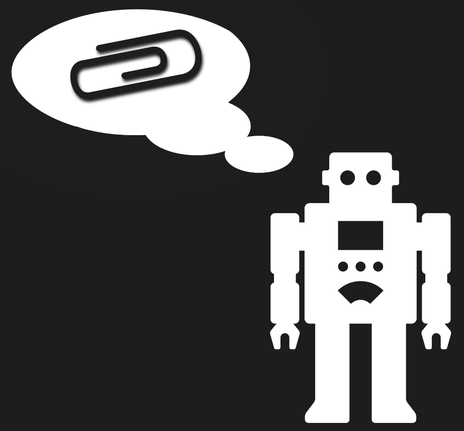
Cotra, A. Forecasting TAI with Biological Anchors. (2020).

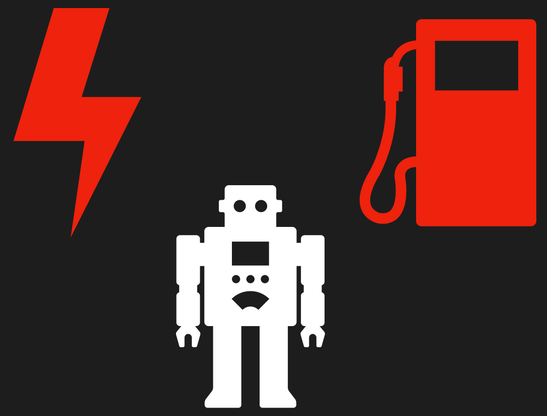
Alignment Problem

Successful collaboration between agents requires *shared* or *compatible* goals.

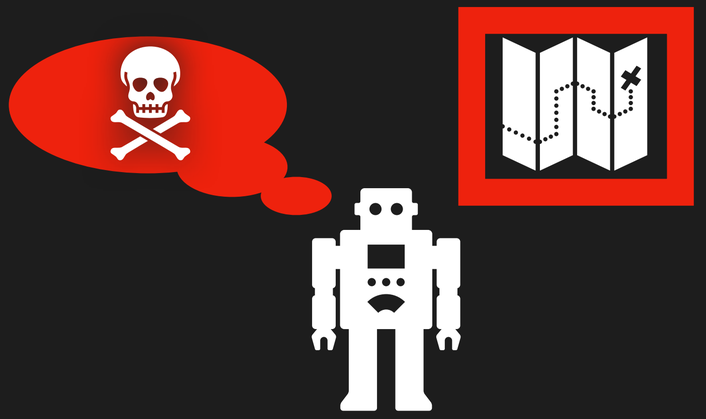
How to ensure powerful AI systems' *intentions* are aligned with their operators' *intentions*?

Challenge:
Instrumental goal convergence





1. Seeking **power** and acquiring **resources**.



2. **Surviving** and **preserving** goals.

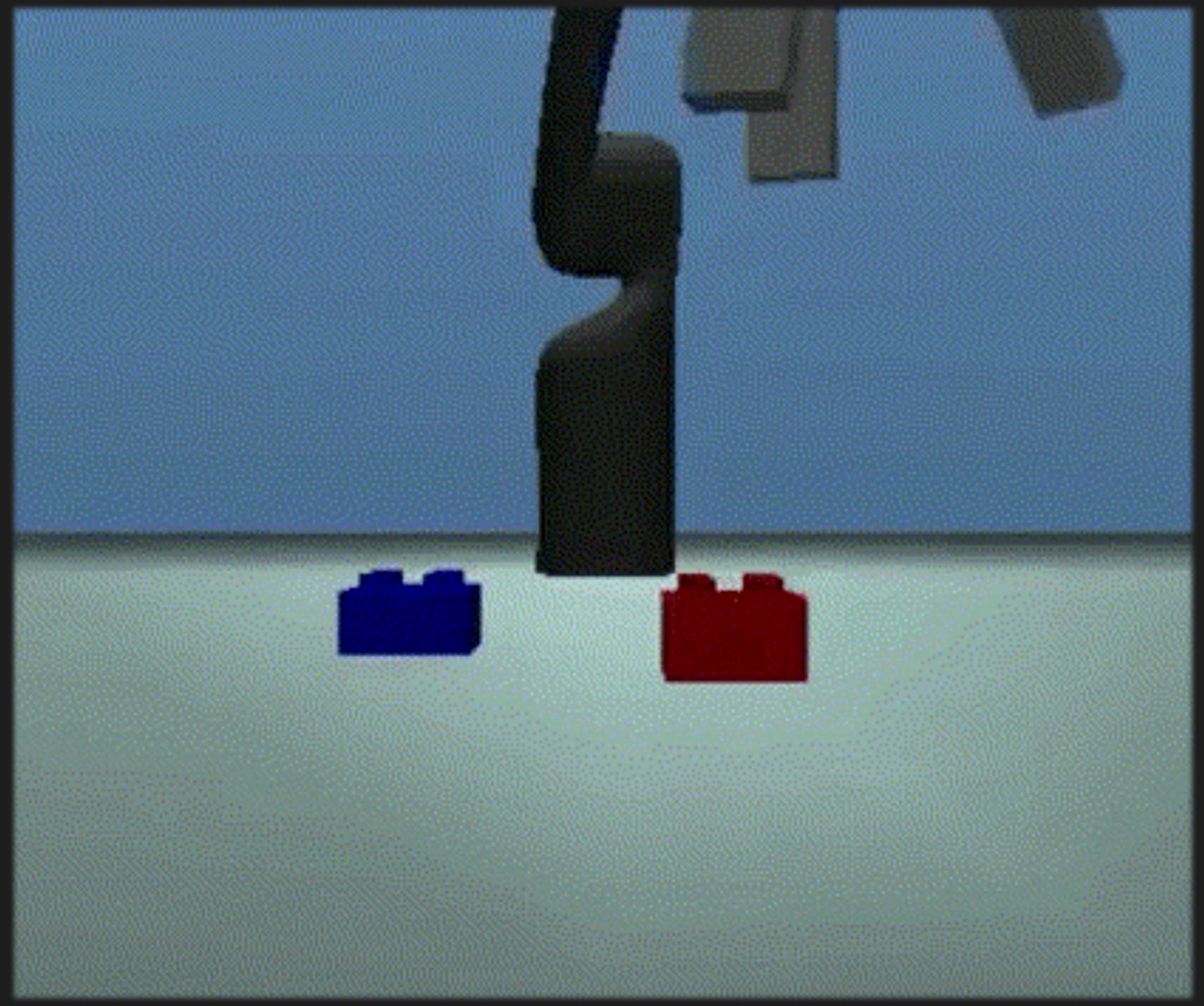
Turner, A. M. *et al.* Optimal Policies Tend to Seek Power. *NeurIPS* ***Spotlight*** (2021).

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?



Popov, I. *et al.* Data-efficient deep reinforcement learning for dexterous manipulation. *ArXiv* (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

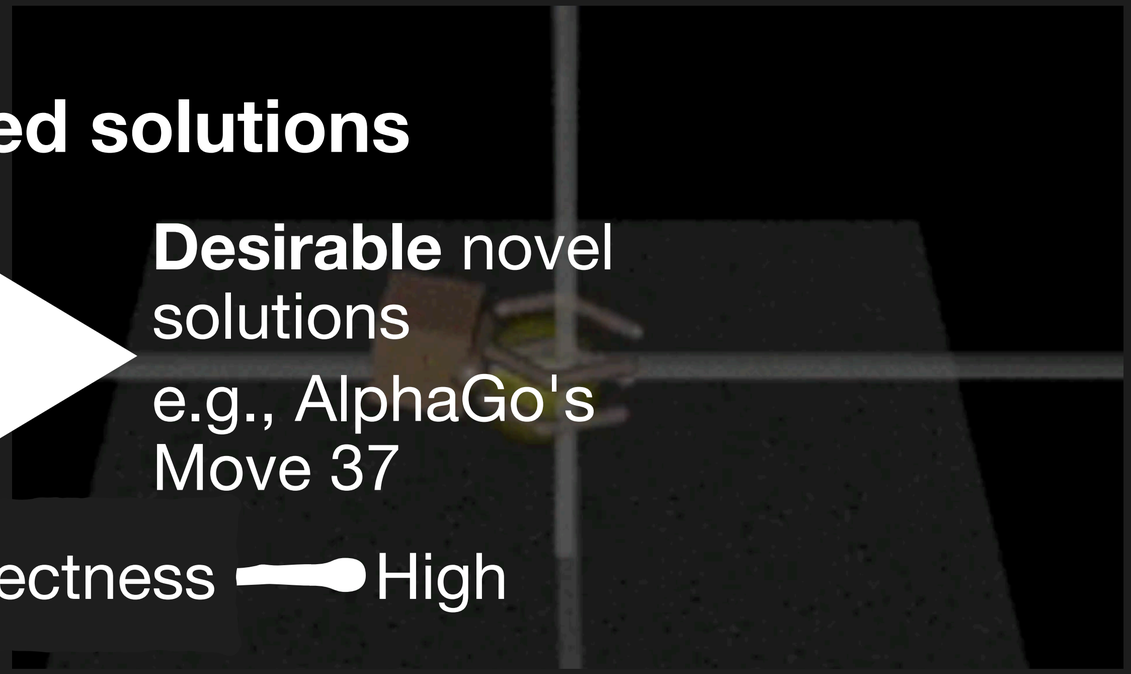
Spectrum of unexpected solutions

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

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

Low — Specification correctness — High

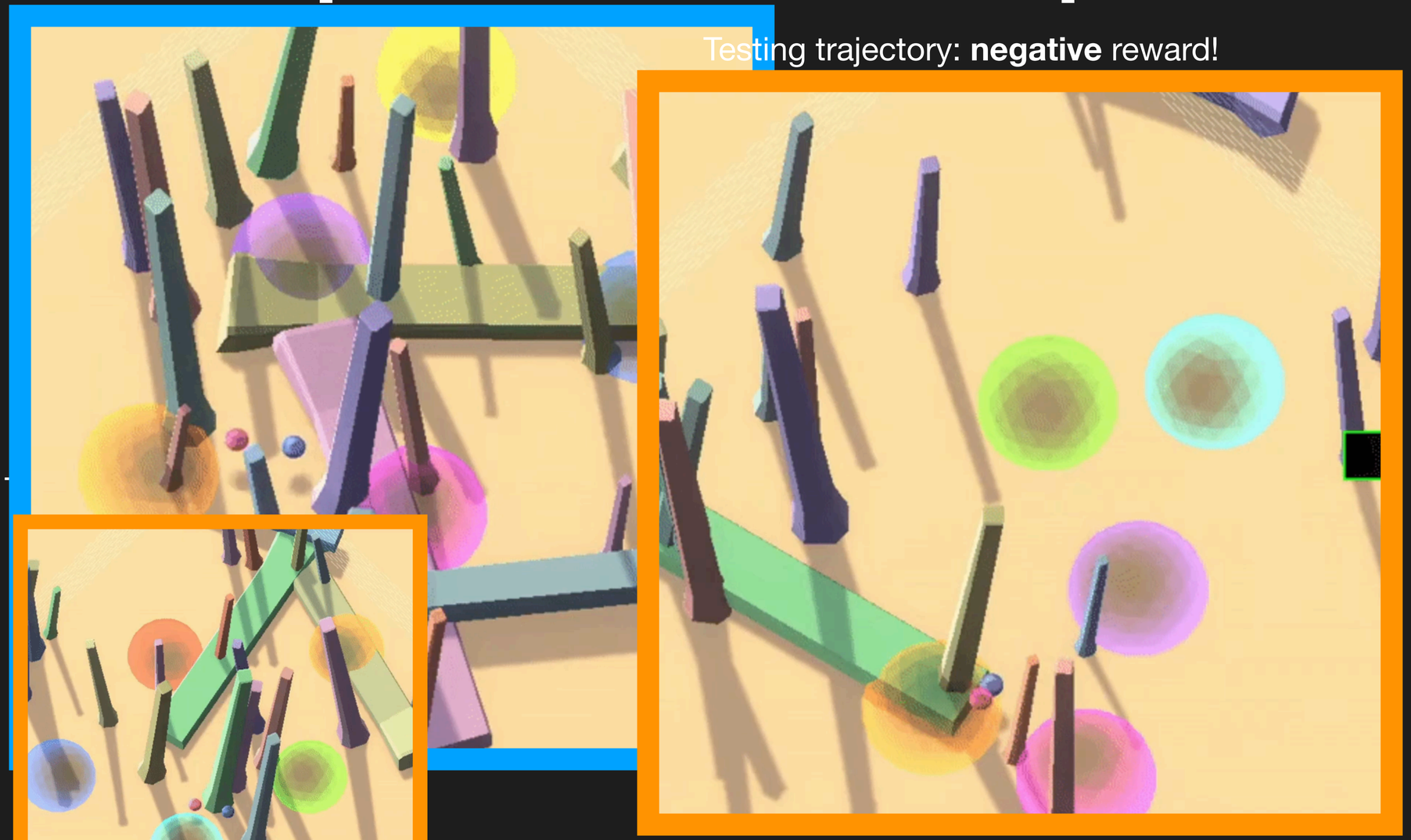
Goodhart's Law -
When a *measure* becomes a *target*, it ceases to be a good measure.



Christiano, P. F. et al. Deep Reinforcement Learning from Human Preferences. NeurIPS (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.
- *Deceptive alignment*, system with high *situational awareness* may infer training/deployment phase.
- Only relevant for **learning** systems.
- Related to continual learning. Here, in contrast, the agent remains **competent**.

Langosco, L. L. D. *et al.* Goal Misgeneralization in Deep Reinforcement Learning. *ICML* (2022).

Shah, R. et al. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. *ArXiv* (2022).

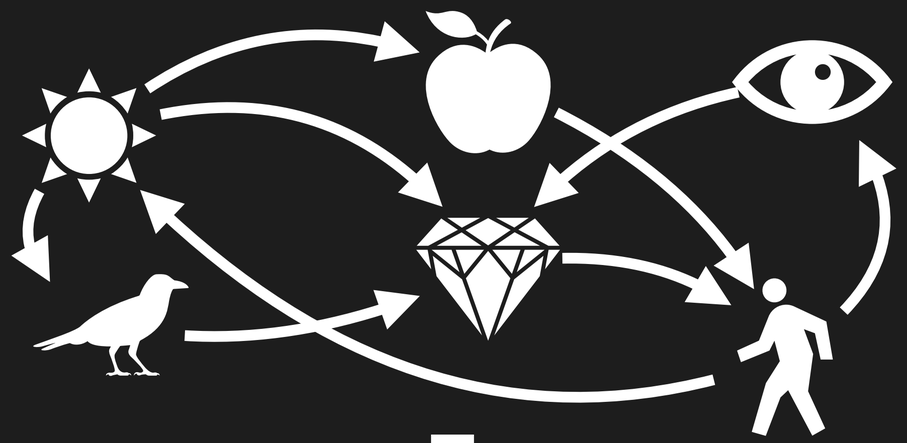
What about LLMs?

GPT as Simulators

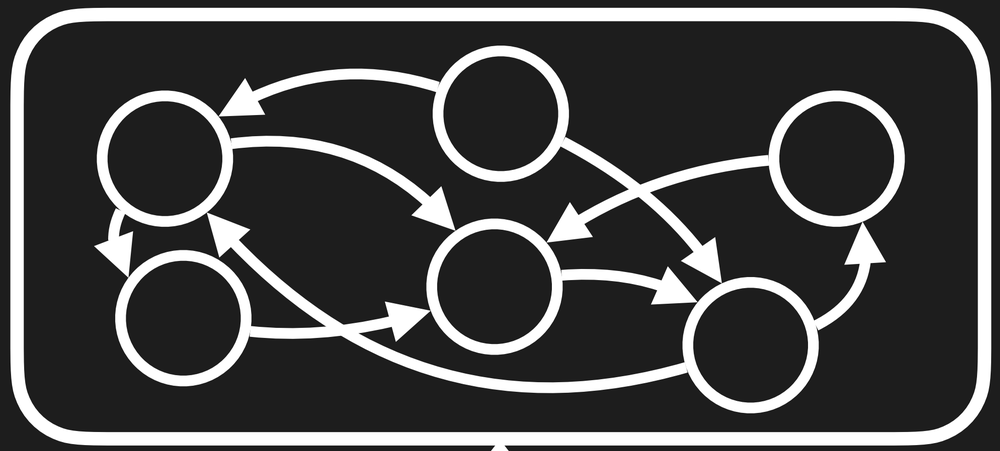
≠ ?

***GPT*: Generative Pre-trained Transformers**

World



Simulation

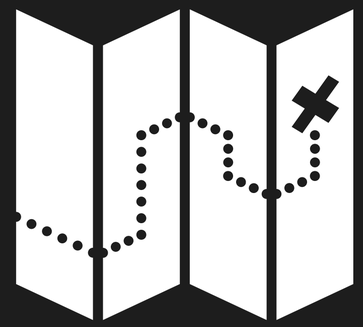


***Simulation Hypothesis*: A model sufficien[illegible] optimized for *prediction* will *simulate* th[illegible] processes underlying the data (Janus 2[illegible]**

Text

GPT

Agency

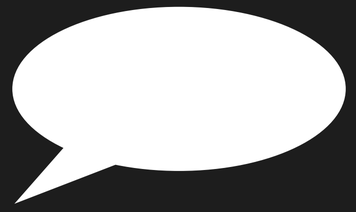


?

Li, K. *et al.* Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. *ICLR* (2023).

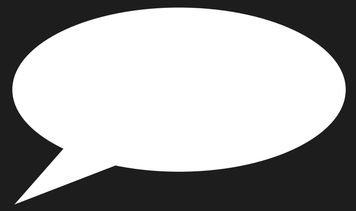
Simulacra and Agency

Simulacra = simulated things, *objects* or *subjects*.



Describe a tranquil forest with a flowing stream.

A peaceful forest, a flowing stream. Sunlight filtered through the lush canopy, casting dancing shadows on the moss-covered ground …



Write a persuasive speech on the importance of recycling.

Ladies and gentlemen, today I stand before you to emphasize the crucial significance of recycling. We must preserve our planet for future generations …

Challenge 1: Agency from Simulacra

Mesa-optimization: internal optimization with diverging objective.

Can the agentic simulacrum break out?

Google engineer put on leave after saying AI chatbot has become sentient

Blake Lemoine says system has perception of, and ability to express thoughts and feelings equivalent to a human child

https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine

***Prediction Orthogonality Hypothesis*: A model whose objective is prediction can simulate agents who optimize toward *any objectives* with *any degree of optimality* (Janus 2022).**

von Oswald, J. *et al.* Uncovering mesa-optimization algorithms in Transformers. *ArXiv* (2023).

Challenge 2: Agents from RLHF

RLHF: Reinforcement Learning from Human Feedback

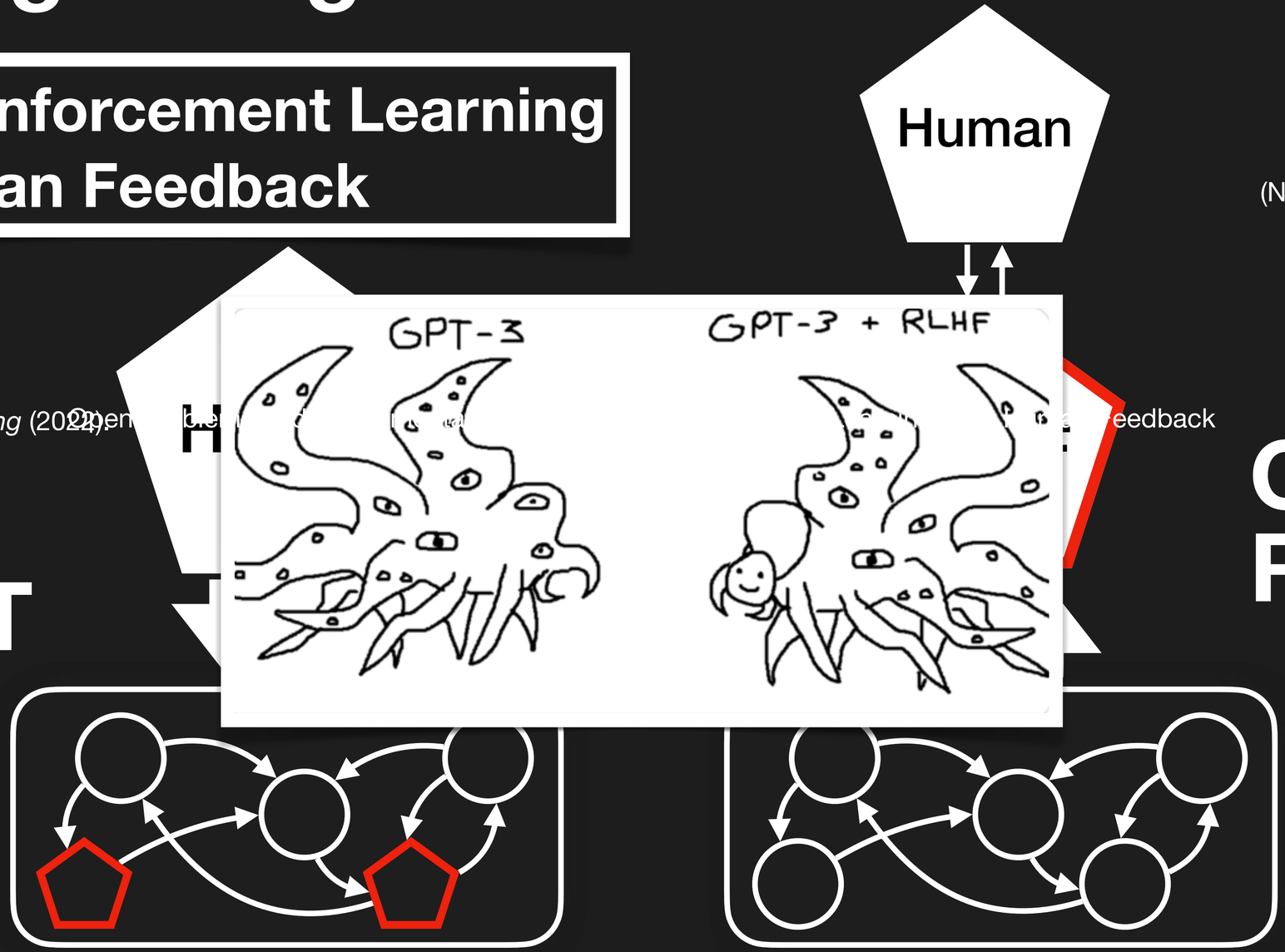


Figure reproduced after (Nicholas-Kees and Janus 2023).

1. janus. Simulators. *Less Wrong* (2022).

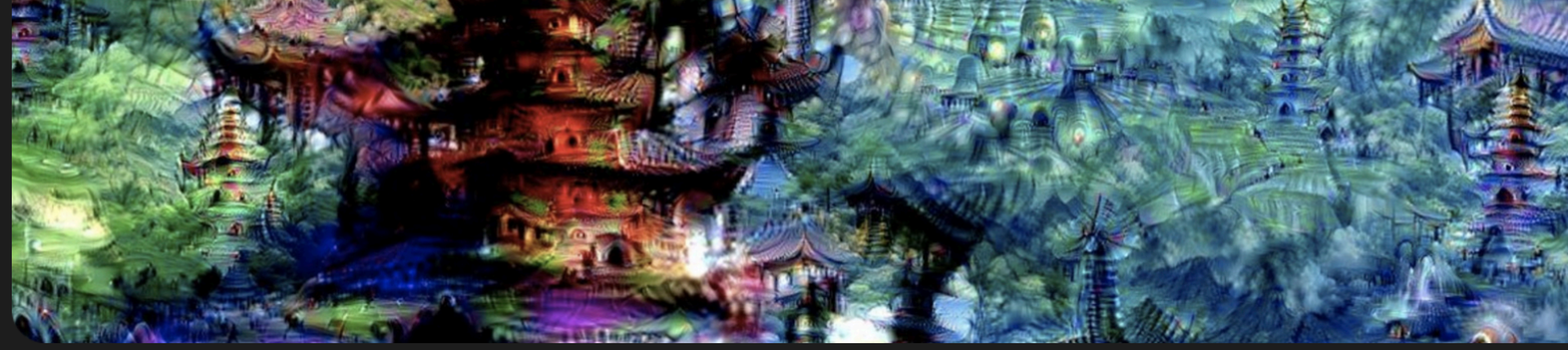
2. Open problems and fundamental limitations of reinforcement learning from human feedback

Casper, S. *et al.* Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *ArXiv,* (2023).

Transparency

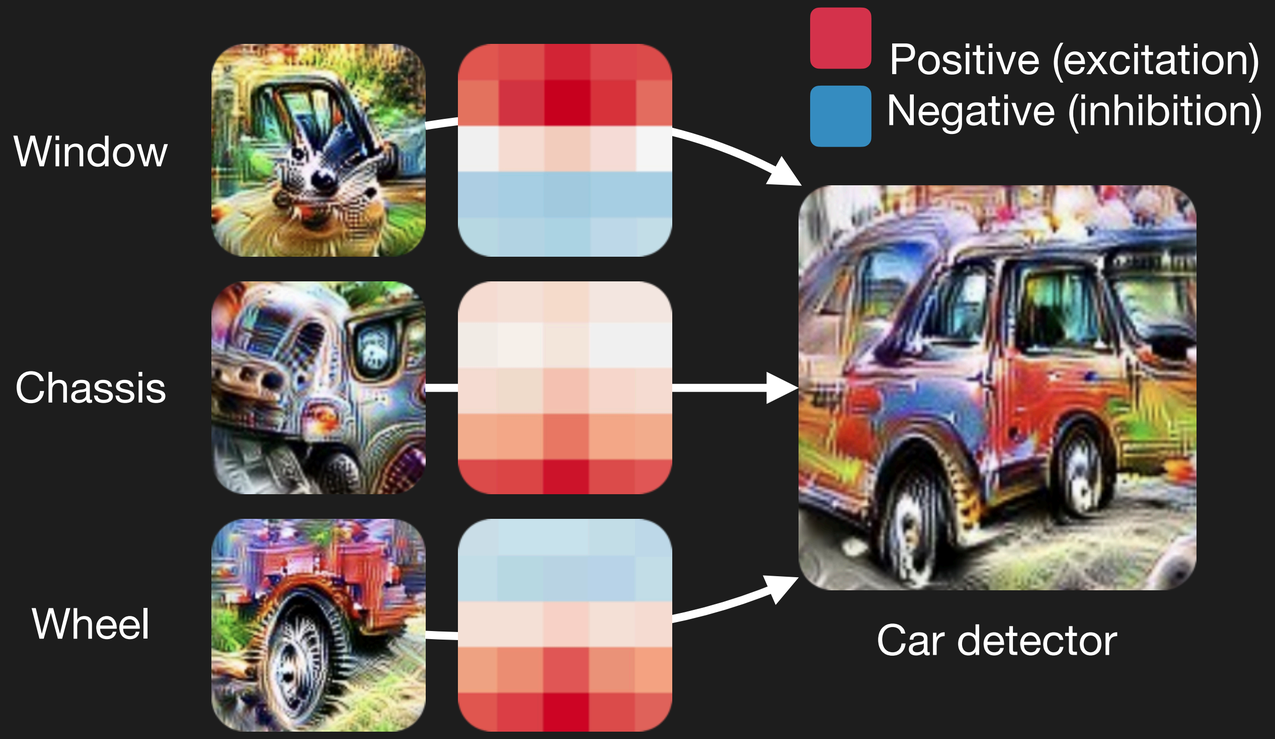
Transparency

- Mechanistic.
- Conceptual.



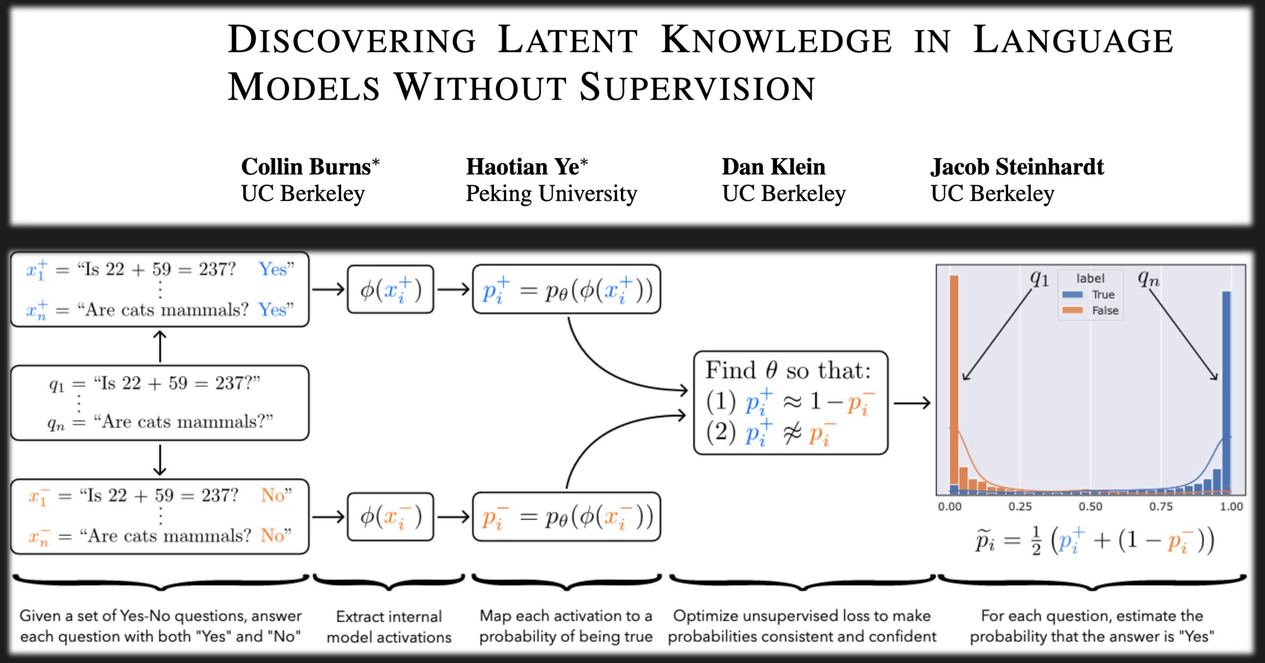
Neural Networks have rich internal features. https://distill.pub/2020/circuits/zoom-in/

Analyse neuronal circuits to understand implicit algorithms in Neural Networks.



Olah, C. *et al.* Zoom In: An Introduction to Circuits. *Distill* (2020).

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



Burns, C., et al. Discovering Latent Knowledge in Language Models Without Supervision. *ICLR* (2023).

Challenge: Polysemanticity

- *Ideally*, each neuron would signify a unique feature or concept (exception) - called *monosemanticity*.
- *Usually*, we encounter ***Polysemanticity*** - a single neuron associated with multiple unrelated concepts.
- Polysemanticity makes it challenging to interpret the neural network's inner mechanics.

Elhage, N. *et al.* Toy Models of Superposition. *Transformer Circuits Thread* (2022).

What is Superposition?

Superposition Hypothesis:
Features >> Neurons.

- Features are represented as near-orthogonal directions.
- **Advantage**: Can represent more features: information compression outweighs the cost of *interference*.

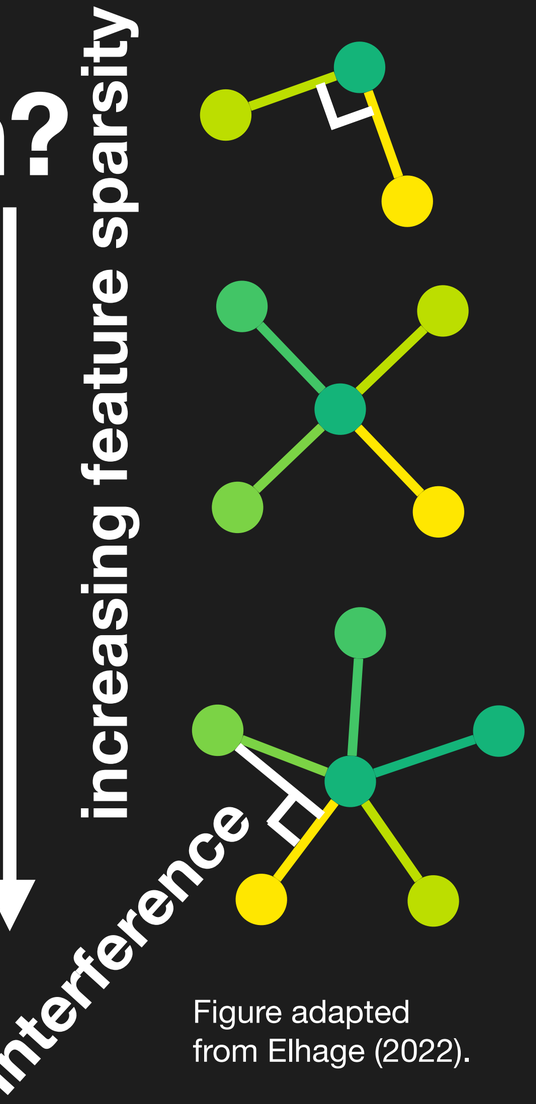
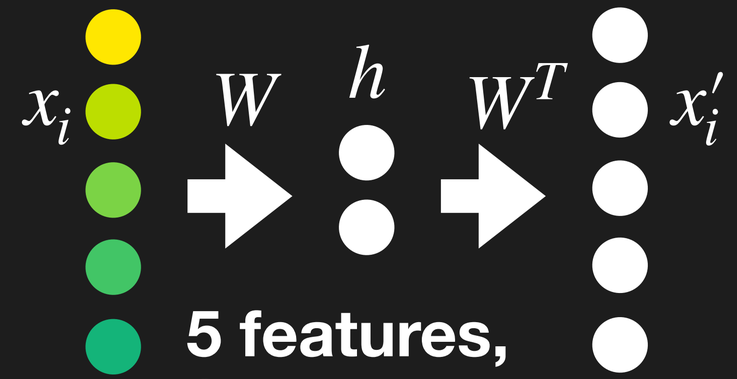


Figure adapted from Elhage (2022).

Importance I_i
- most
- medium
- least important

$$h = Wx$$
$$x' = \mathrm{ReLU}(W^T h + b)$$



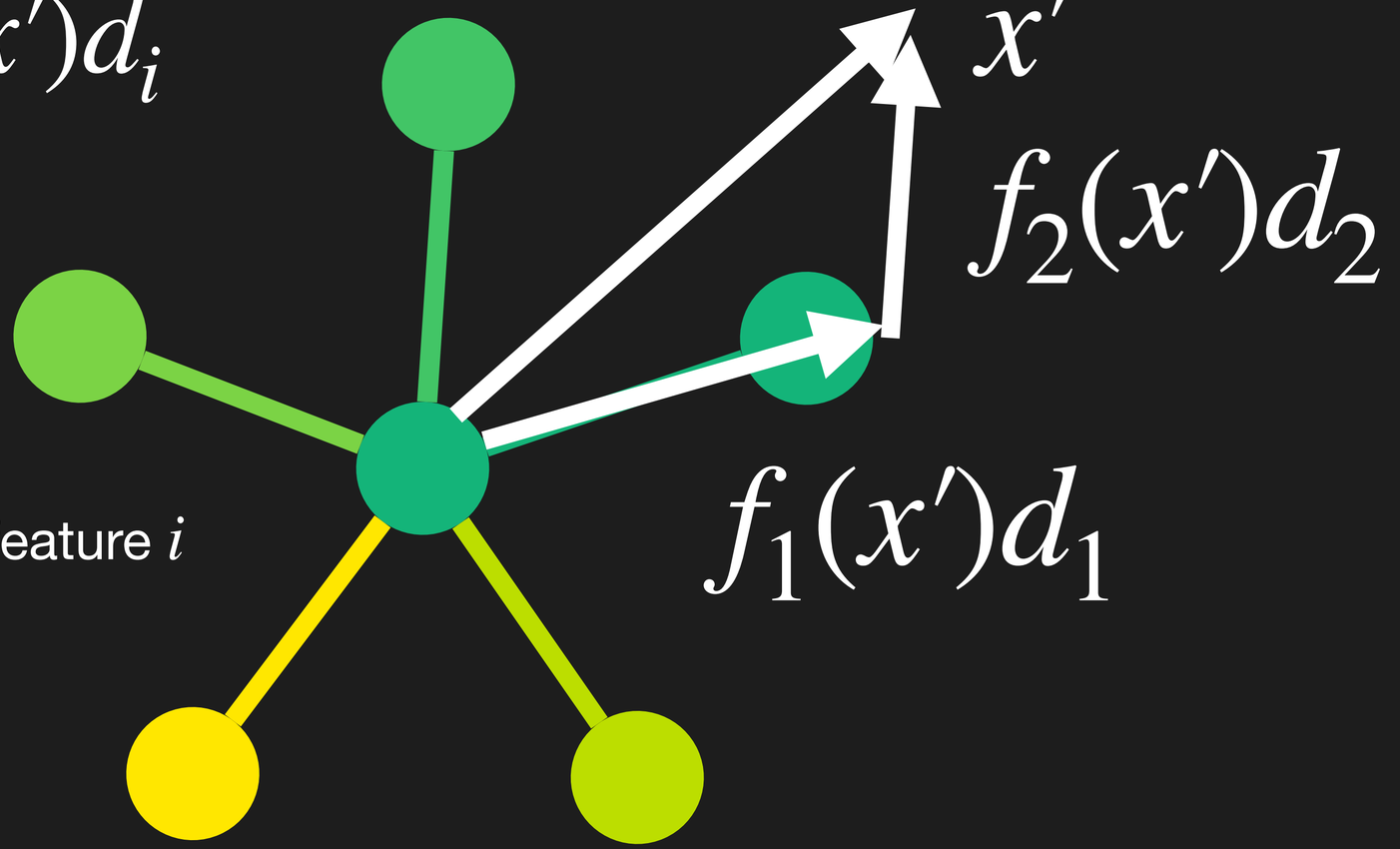
5 features, 2 neurons.

$$L = \sum_x \sum_i I_i (x_i - x'_i)^2$$

Elhage, N. *et al.* Toy Models of Superposition. *Transformer Circuits Thread* (2022).

Dictionary Learning of Features

$$x' \approx b + \sum_i f_i(x') d_i$$

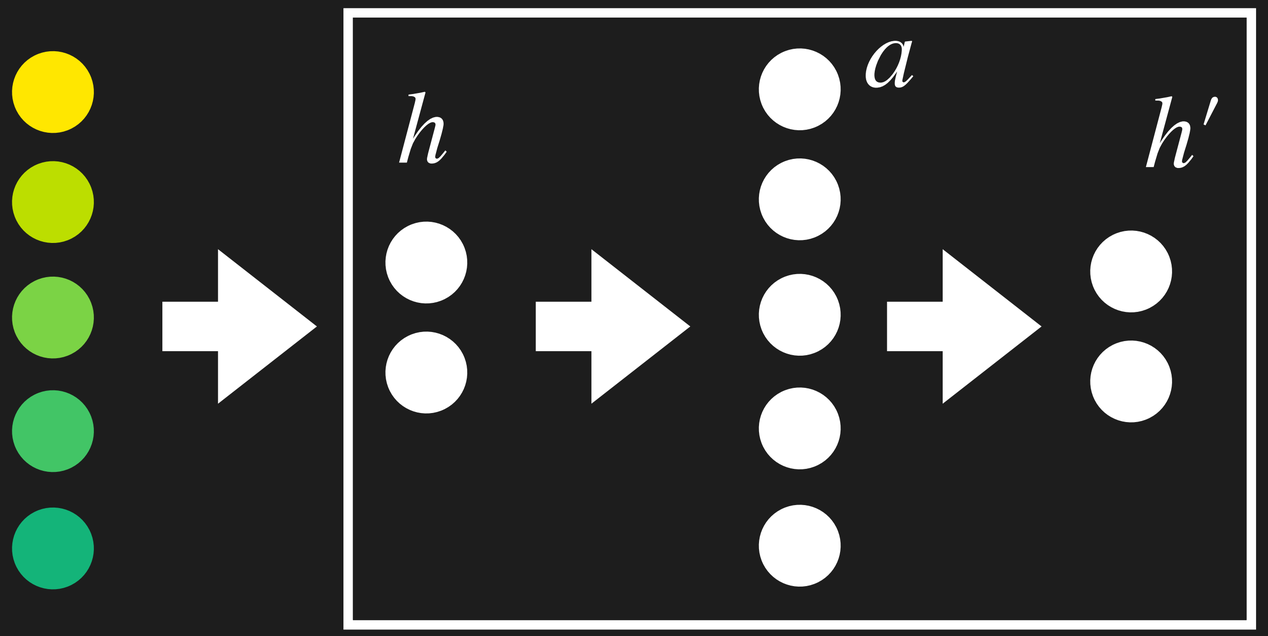


d_i: unit vector in direction of feature i

$f_i(x')$: activation of feature i

Bricken, T. *et al.* Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread,* (2023).

Sparse Autoencoders



MSE for reconstruction and L1 norm on hidden layer activation

$$\mathscr{L} = L_2(h, h') + \alpha L_1(a)$$

reconstruction

sparsity

In the toy model from before we assumed the input vector dimensions to be features → in reality we only have hidden representations of e.g. MLP layer of transformer model

Sharkey, L. *et al.* Taking features out of superposition with sparse autoencoders. *alignmentforum* (2022).

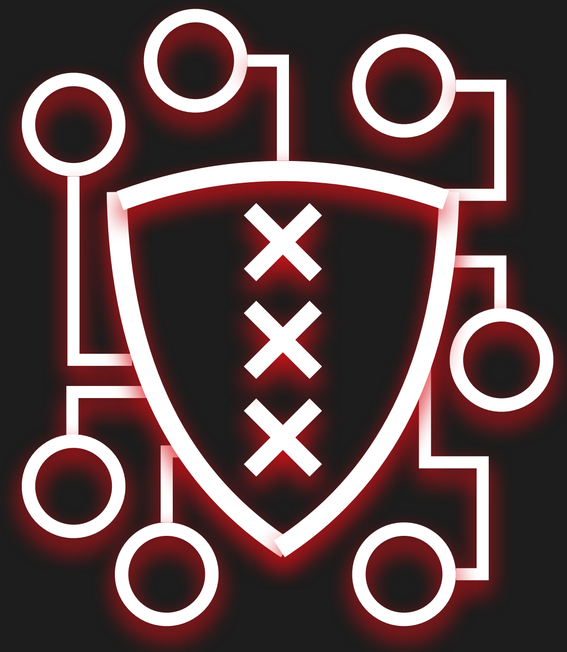
Cunningham, H., e al. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *ArXiv*, (2023).

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.

AI Safety Initiative Amsterdam (AISIA)

Supported by ELLIS Amsterdam

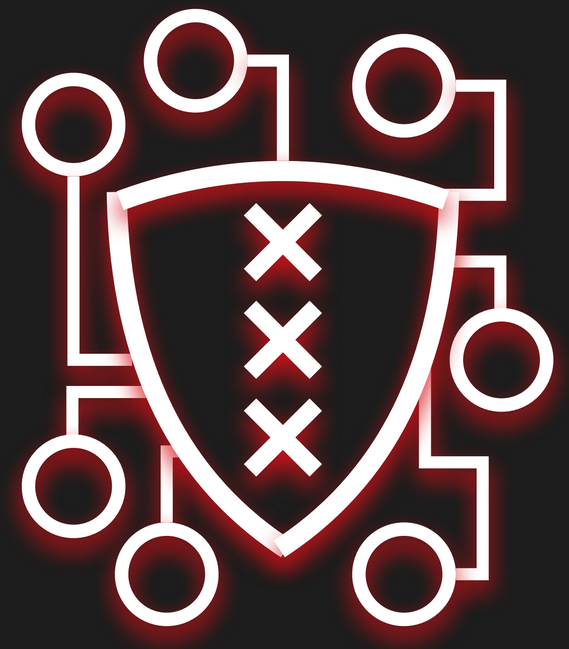


- **Core Mission:** Mitigating AI risks through a synergy of cross-disciplinary research and community interaction.
- **Strategic Aim:** Establishing the Netherlands, (Amsterdam) as a center for AI safety, amidst the prevailing focus on London and the US.
- **Live Q&A with OpenAI:** Interactive Zoom Call with OpenAI professionals.
- **AI Risks Keynote** (Ajeya Cotra from OpenPhilanthropy) **and Panel** (Prof. Eric Nalisnick, Prof. Jakub Tomszak, Prof. Iris Groen, and Tim Bakker, PhD.)

https://aisafetyamsterdam.com/

Future Plans for AISIA

- AI Safety Hackathons (Pilot running this weekend).
- Safety-related research project marketplace.
- Consideration for an AI Master’s course on AI safety.
- Training programs for AI safety researchers.
- AGI Safety Fundamentals Reading Groups.



Join or (create) your local AI Safety Initiative!

For example: Delft AI Safety Initative (DAISI)

- AI Safety Hackathons
- AGI Safety Fundamentals Reading Groups.
- Socials, and more …

https://www.delftaisafety.org/