

Taming Simulators


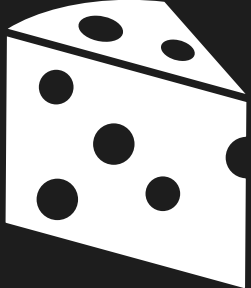
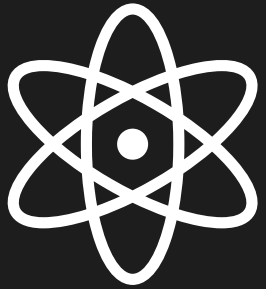
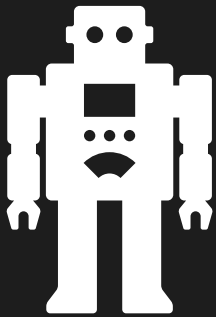
Alignment of Large Language Models

Leonard Bereska and Efstratios Gavves

AAAI Symposium 2023 Singapore - Human-AI Collaboration.

Leonard Bereska, July 17th, 2023.

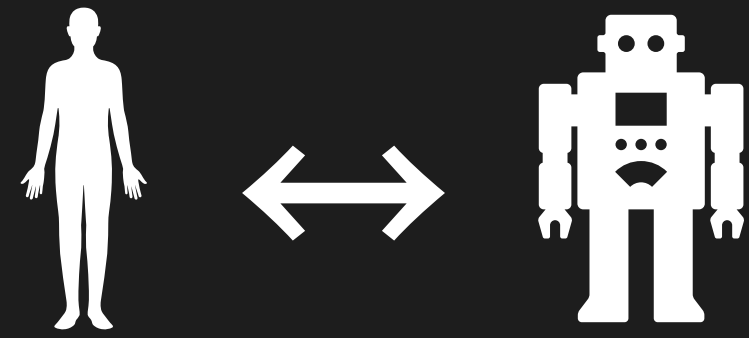
Hey, I'm Leo!

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



Human-AI Collaboration → Alignment Problem

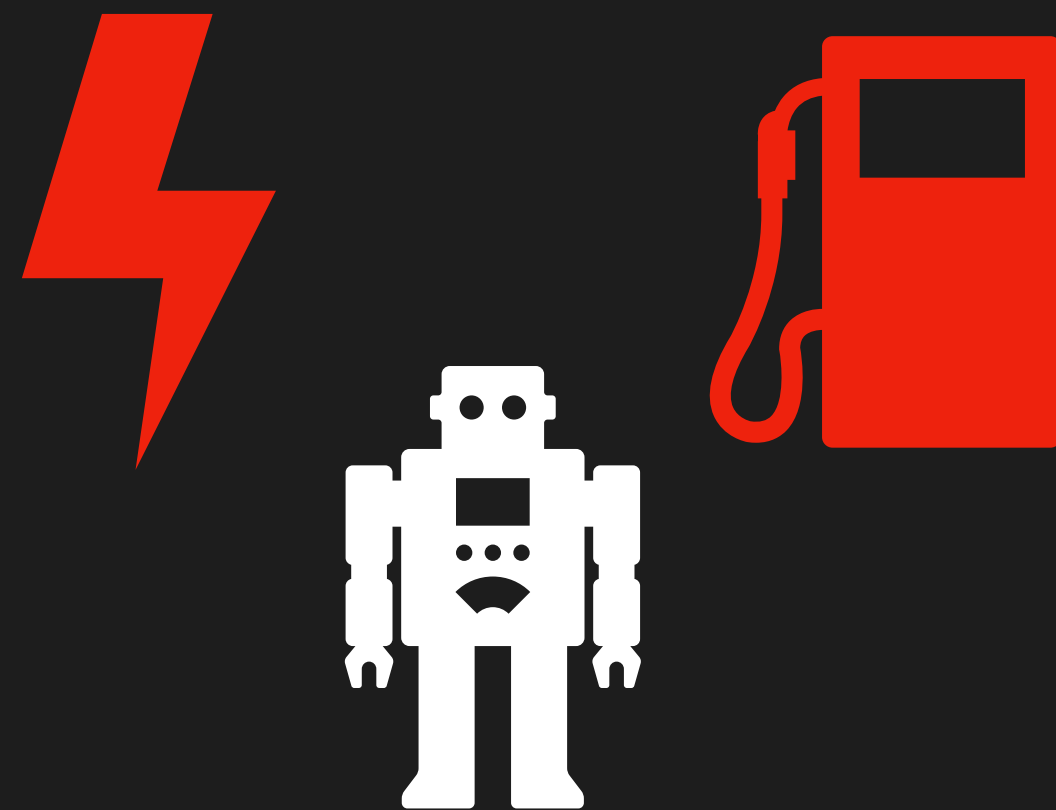
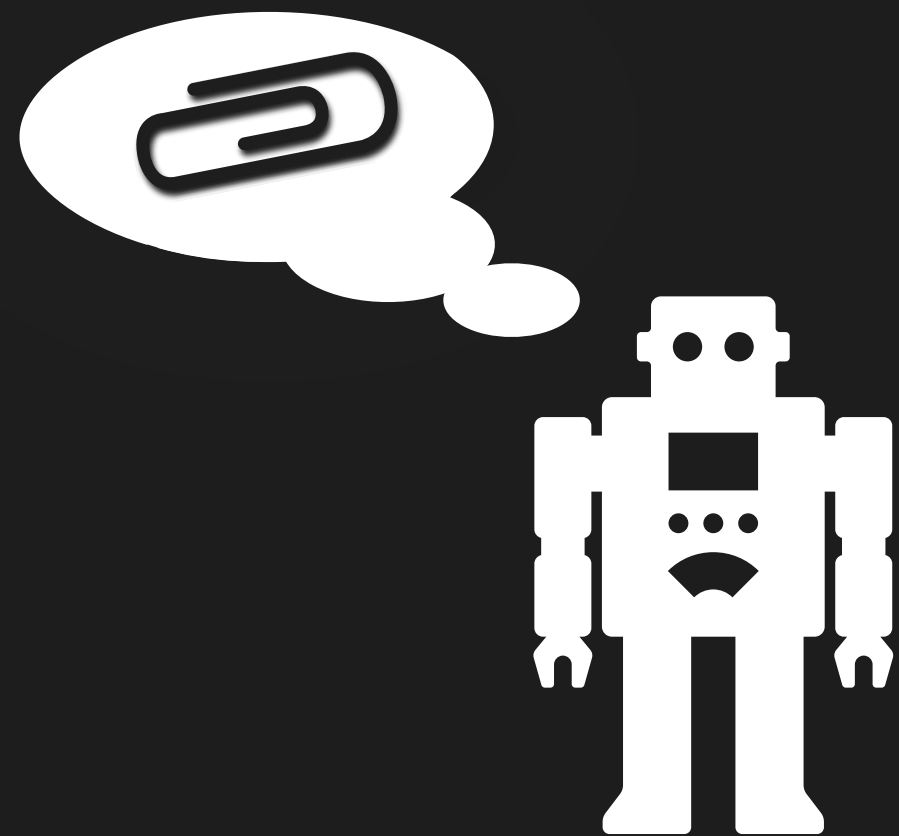
Successful collaboration between agents requires *shared* or *compatible* goals.



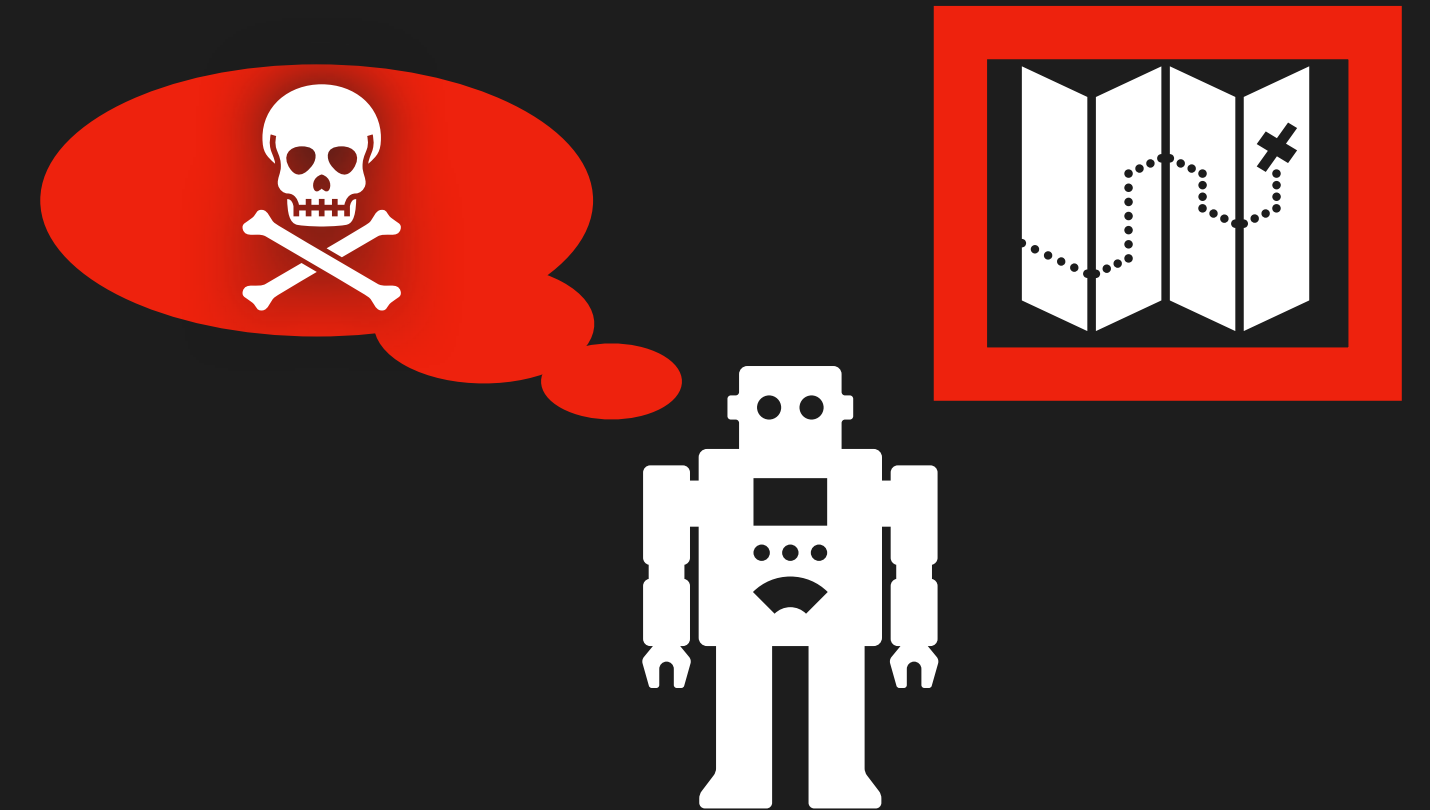
How to ensure powerful AI systems' *intentions* are aligned with their operators' *intentions*?

Challenge:

Instrumental goal convergence

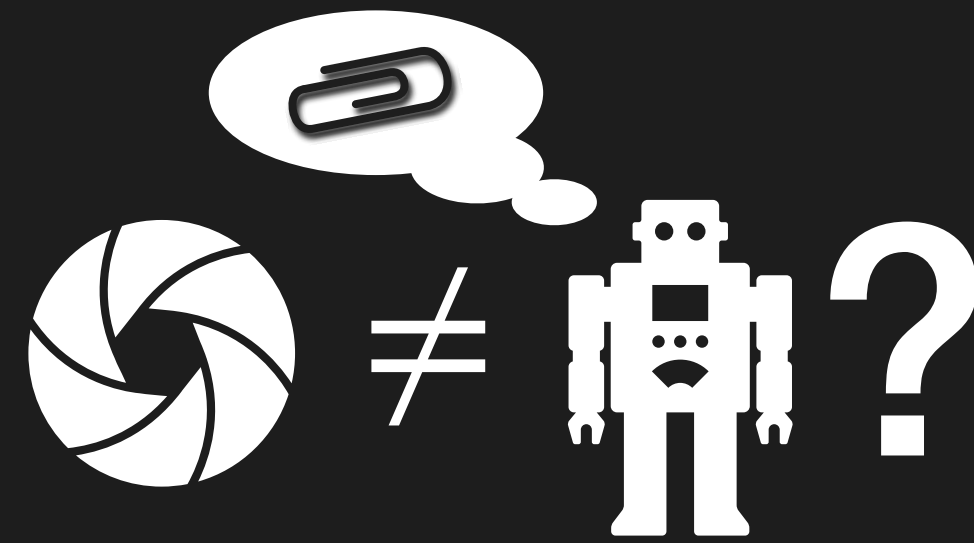


1. Seeking **power** and acquiring **resources**.



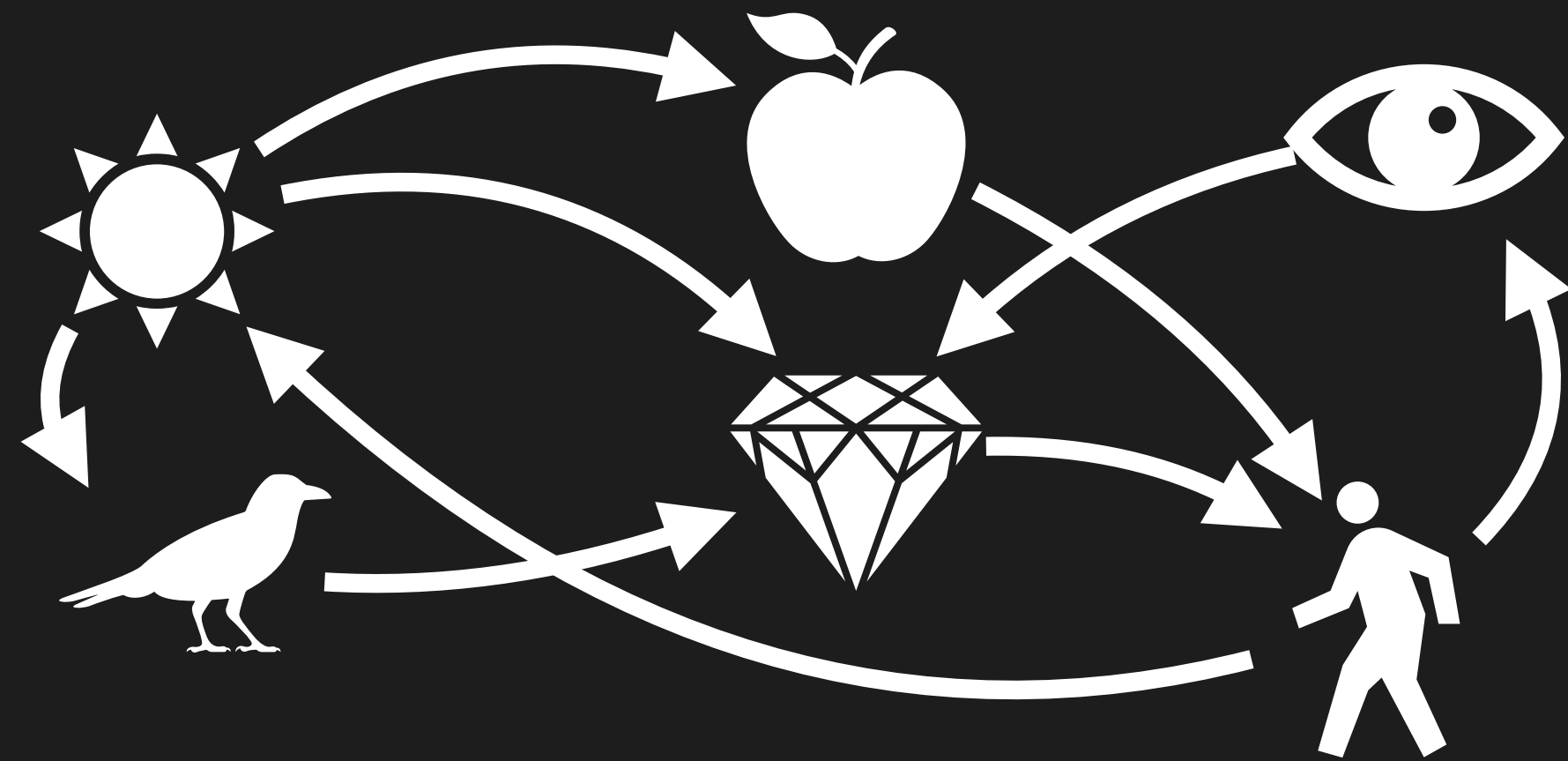
2. **Surviving** and **preserving** goals.

GPT as Simulators

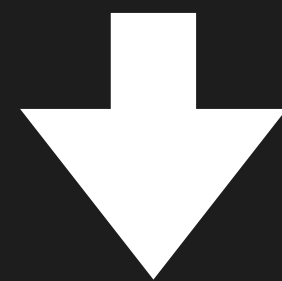
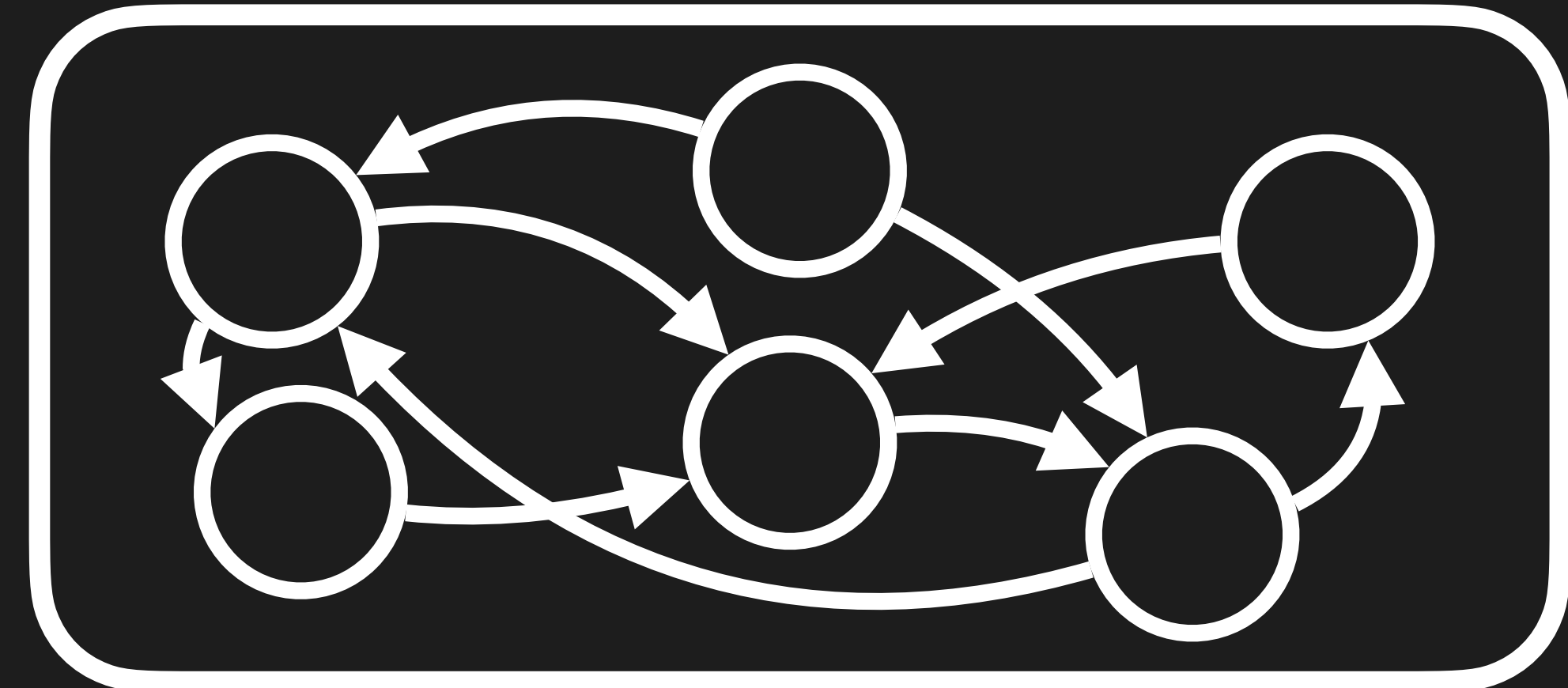


GPT: Generative Pre-trained Transformers

World



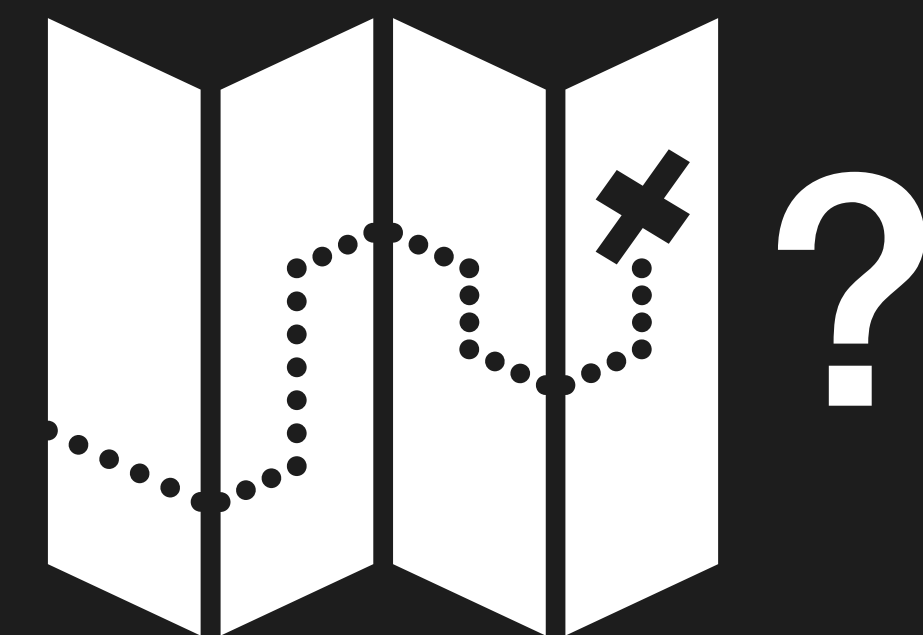
Simulation



Simulation Hypothesis: A model sufficiently optimized for *prediction* will *simulate* the processes underlying the data (Janus 2019)



Agency

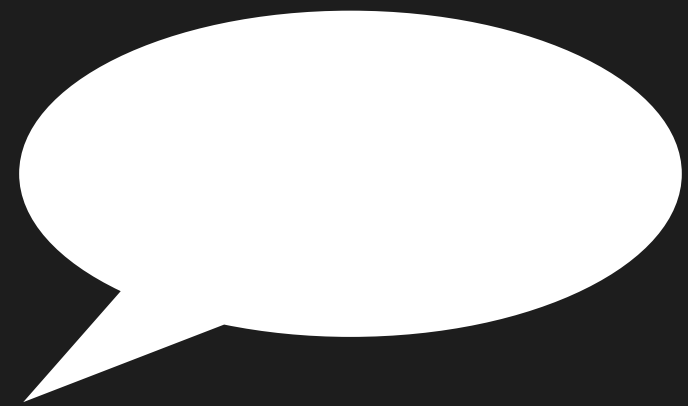


Text

GPT

Simulacra and Agency

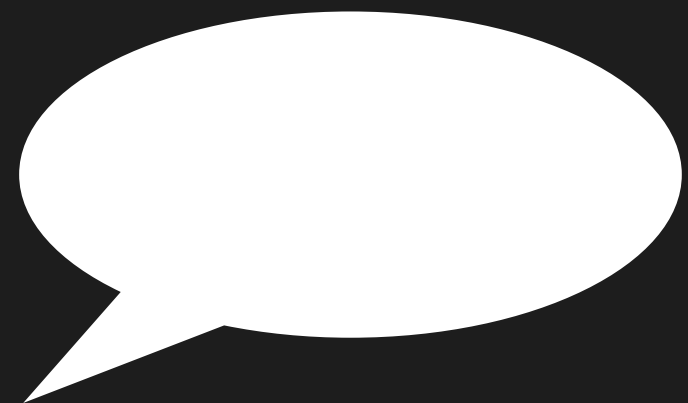
Simulacra = simulated things, objects or **subjects**.



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 ...

Challenges

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

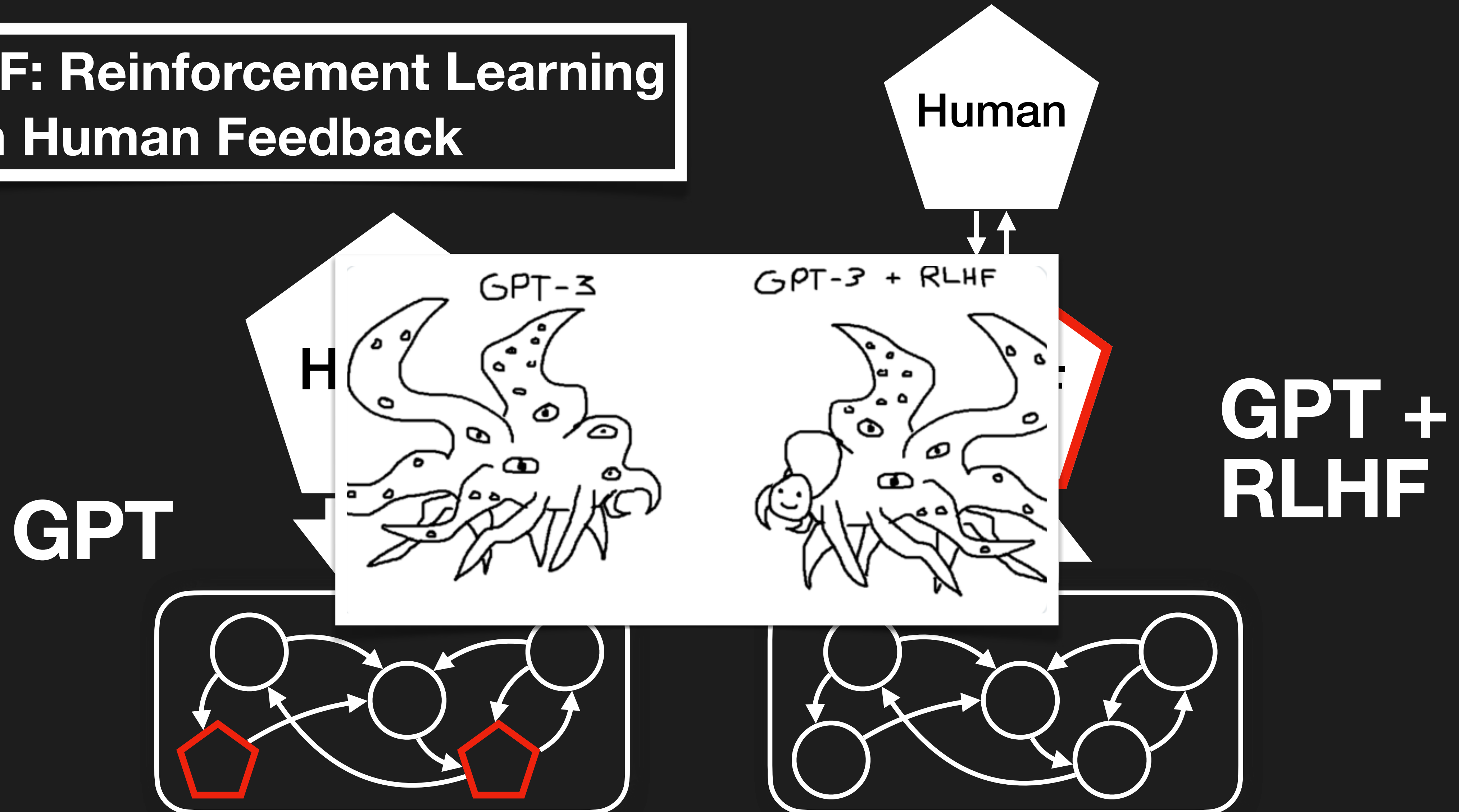
Blake Lemoine says system has perception of, and ability to express thoughts and feelings equivalent to a human child

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



Challenge 2: Agents from RLHF

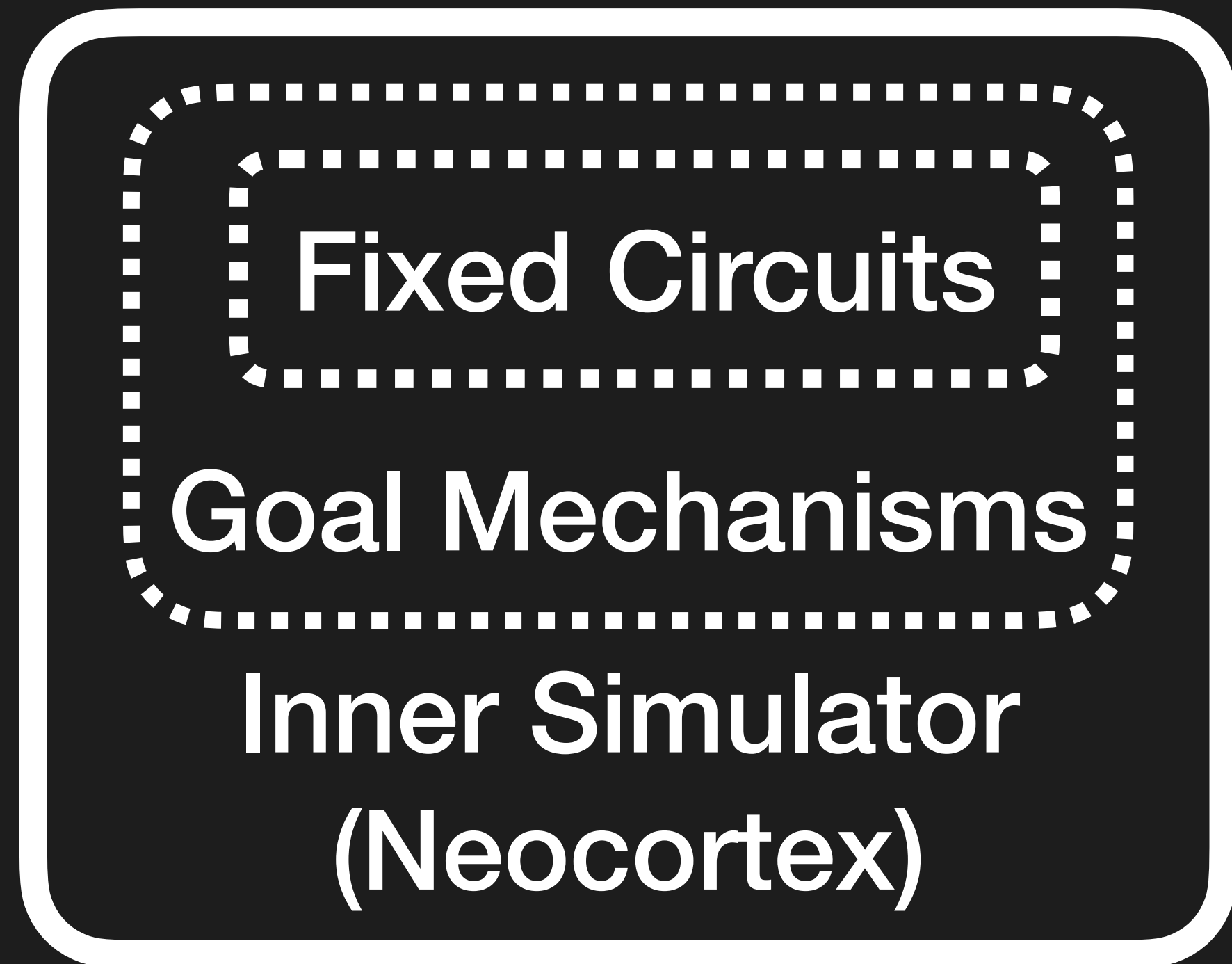
RLHF: Reinforcement Learning from Human Feedback



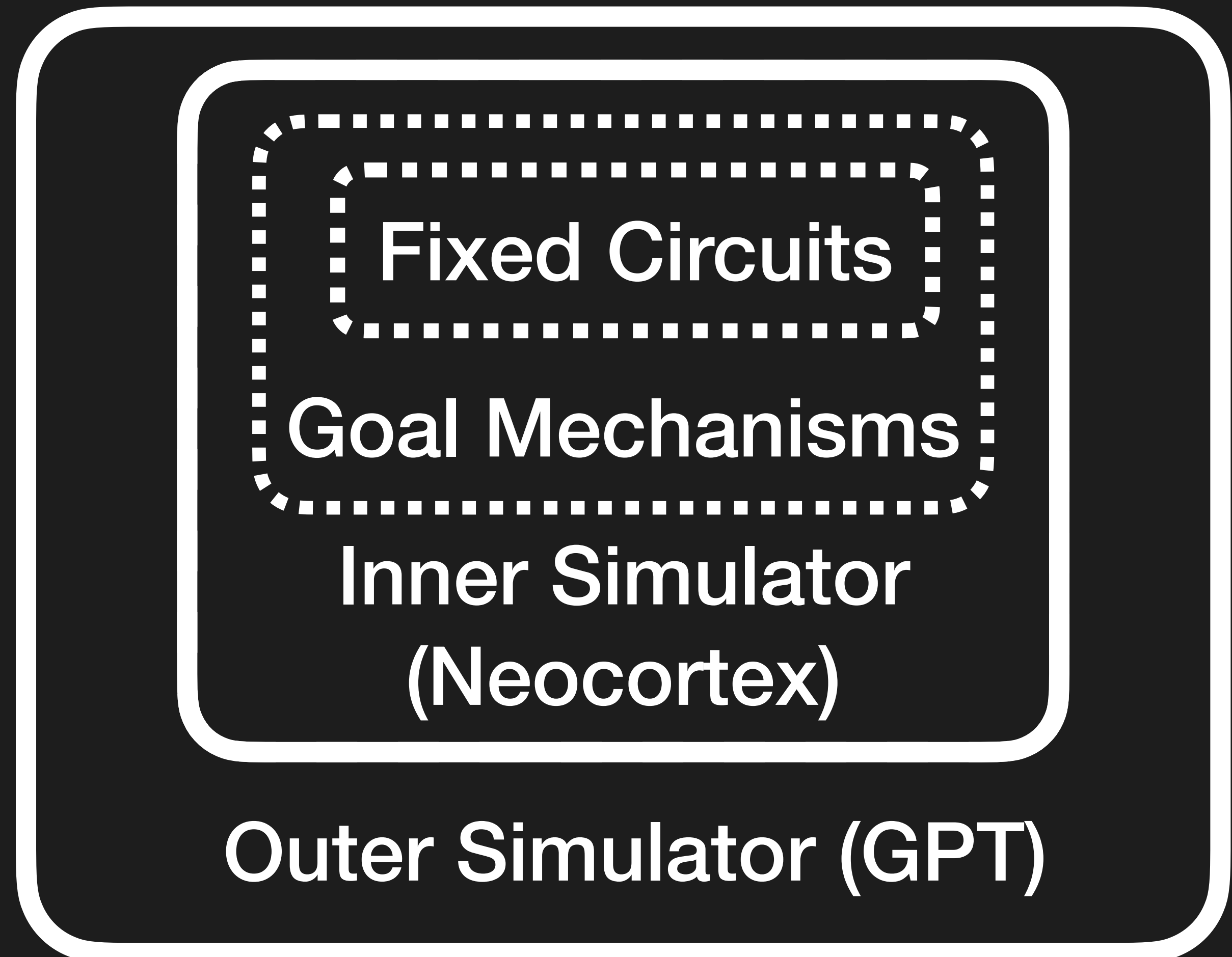
Visions

Vision 1: Cyborgism

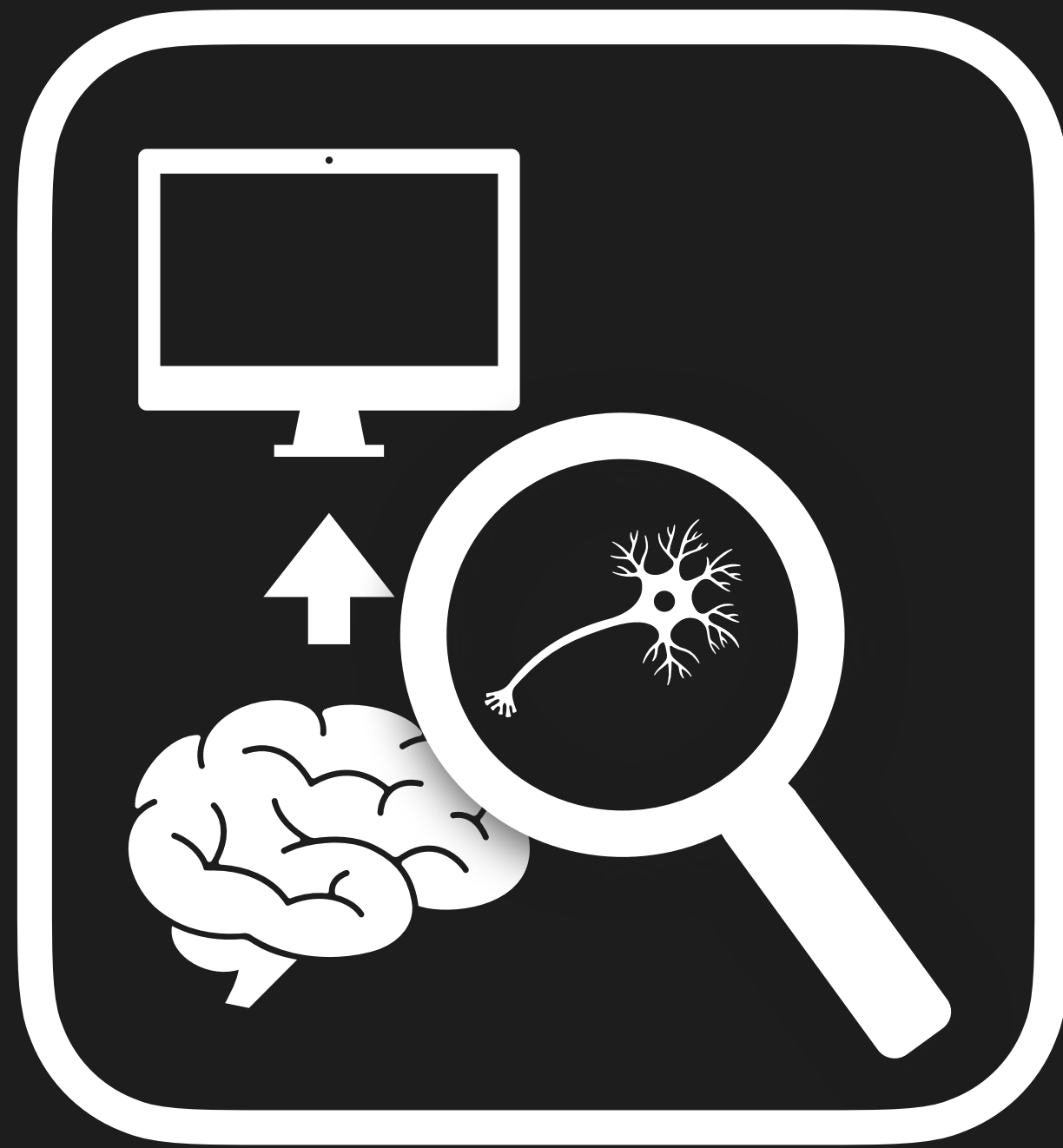
Human



Human + GPT



Vision 2: Emulation



Whole brain
emulation



Cognitive
emulation



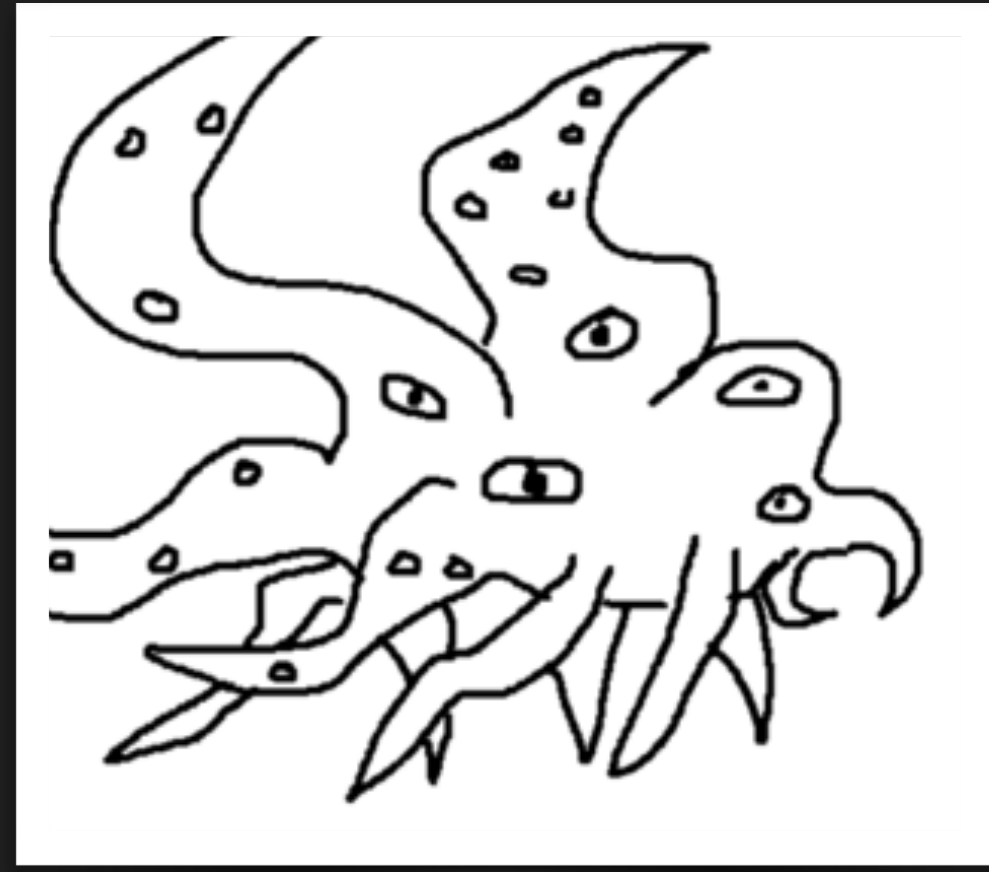
Paperclip
maximizer

Conclusion

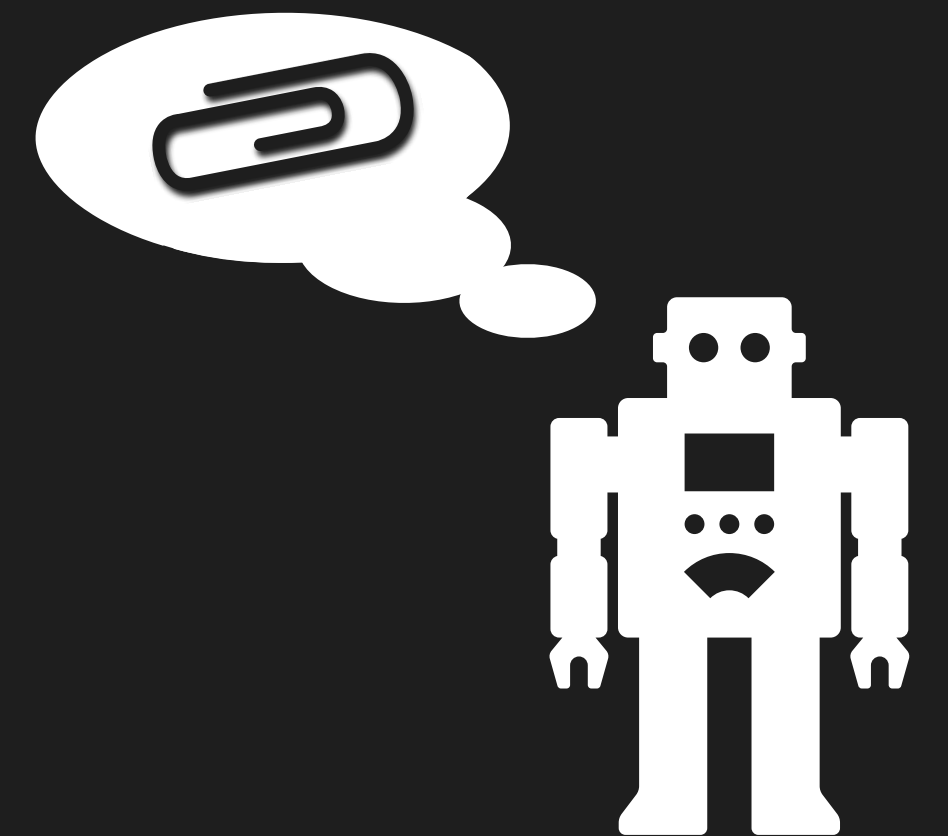
- We may develop more-powerful-than-human AI in the foreseeable future.
- Powerful AI may not be beneficial by default.
- Continuing on the current path holds the potential for catastrophic outcomes.
- More research necessary to align powerful AI with humanity's existence.

***A pessimist sees the difficulty in every opportunity;
an optimist sees the opportunity in every difficulty.***

Winston Churchill



**Thank you for your
attention!**



Questions for Human-AI Collaboration

- **What are challenges and opportunities in designing AI systems that can collaborate with humans in a natural way?**
 - How can we trust AI systems when we optimize for behavior and can't monitor intentions?
 - How can we guard against emotional manipulation by AI systems?
- **What are potential connections and synergies between human-AI and human-human collaboration?**
 - How can we build empathy and compassion or moral reasoning into AI systems?
 - Can we learn from study of psychopaths as a model for misaligned AI systems?