

Societal Impacts and Agent Safety

Prithviraj Ammanabrolu

Special thanks to DeepMind AGI Safety Team for their input



A blue robot with a rectangular head and a visor is reading a book. The book cover is dark with the text 'Shakespeare Typed by Monkeys' in a cursive font and 'Volume 78' below it. The robot is in a futuristic, dimly lit environment with circular portholes in the background. A semi-transparent black box with a cyan border is overlaid on the right side of the image, containing text.

What is an agent?

WHATEVER YOU WANT
IT TO BE

A blue robot with a large, rectangular visor and a grid-patterned mouth is reading a book. The book's cover is dark with the text 'Shakespeare Typed by Monkeys' in a cursive font and 'Volume 78' below it. The robot is in a futuristic, dimly lit environment with circular portholes in the background. A dark semi-transparent box on the right contains white text.

What is an agent?

JK AGENTS ARE A
COMBO OF:


GROUNDING, AGENCY,
MEMORY, REASONING,
LEARNING

SIMPLEST VERSION: ACT
WITH TOOLS IN A LOOP
TO SOLVE PROBLEMS

A blue robot with a large, rectangular head and a wide, open mouth is reading a book. The book's cover is dark with the text 'Shakespeare Typed by Monkeys' in a cursive font and 'Volume 78' in a sans-serif font. The robot is in a futuristic, dimly lit environment with circular portholes in the background. A dark rectangular box with a light blue border is overlaid on the right side of the image, containing text.


What are reasoning capabilities?

ALSO WHATEVER YOU WANT IT TO BE

A blue robot with a rectangular head and a visor is reading a book. The book cover is dark with the text 'Shakespeare Typed by Monkeys Volume 78' in a stylized font. The robot is in a futuristic environment with circular windows and a glowing blue screen in the background.

What are reasoning capabilities?

SOLVING PROBLEMS
THAT REQUIRE
SEQUENTIAL DECISION
MAKING OVER MANY
STEPS

A blue robot with a large, rectangular visor and a grid-like mouth is reading a book. The book cover is dark with white and red text that reads "Shakespeare Typed by Monkeys Volube 78". The robot is in a control room with various screens and panels in the background. A semi-transparent dark box with a light blue border is overlaid on the right side of the image, containing text.

What is safety?

WE ACTUALLY ~KNOW
THIS ONE

CYBERSECURITY, E.G.
PREVENTING PROMPT
INJECTIONS

ALL THE WAY TO X-RISK

PREVENTING
UNDESIRABLE OUTPUTS

Not Safety vs Capability, Safety AND Capability



Safety

If you develop a safety method that trades off large capabilities, no one will use it.

Scales to overseeing, say, O(100m) requests

It is a spectrum!!

Capability

If you develop a very risky exploration method that improves capabilities, it is also unlikely to be adopted.

Scales with more data and FLOPS

What scalable safety-capability levers do we have?



Model Priors

Mid-training Data
Supervised Finetuning Data



Rewards

Inference-time compute for rewards



Environments

Task Diversity (multi-objective)
Task Complexity (multi-agent, horizon length...)



Policy Training

Parameter size
Step count + Inference-time compute for rollouts

What happens when we have advanced AI systems?

- We are on a path to AI Systems that will outperform most humans at most cognitive tasks by the end of the decade
- What does this mean for all the humans?

AI Safety, the opposing viewpoints

There will be
no issues

AI will kill us
all

A secret
third thing

Value Alignment

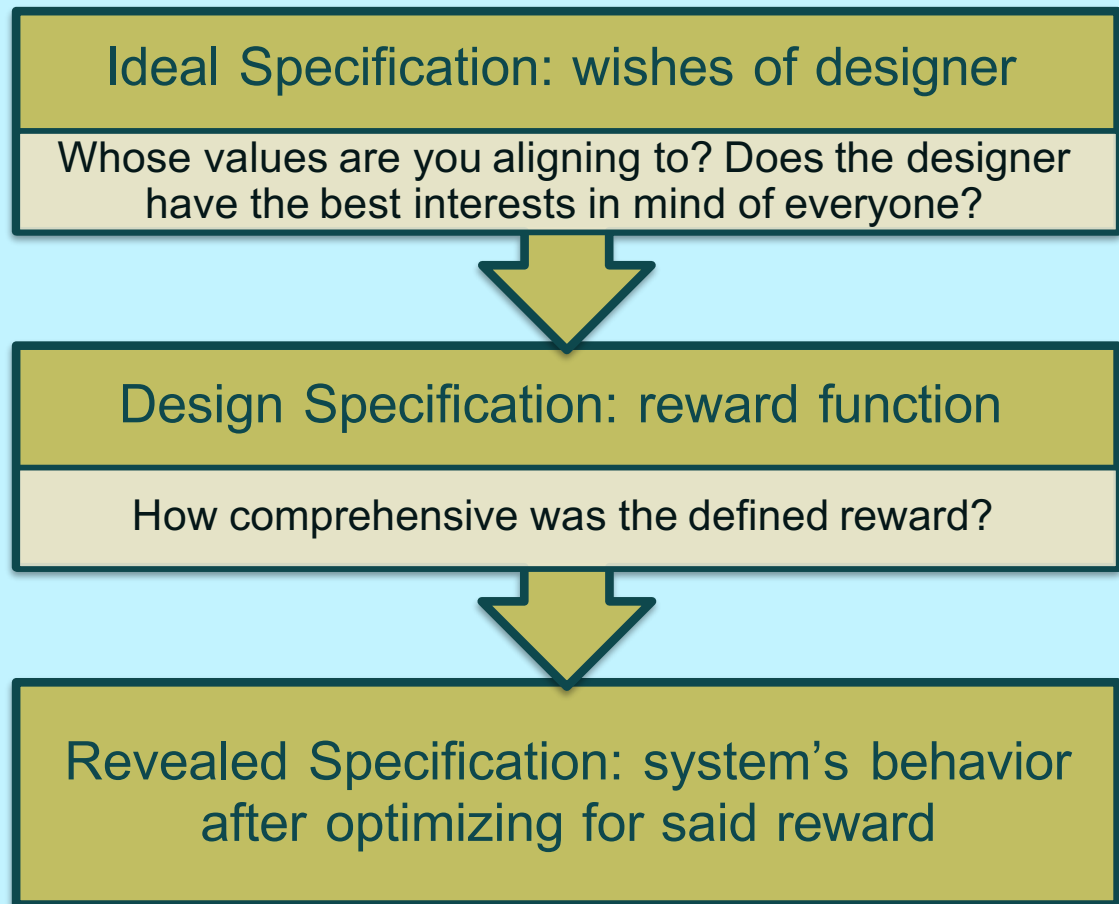


Many definitions



Simplest one (paraphrased): actions that an agent performs in a state are measurably close to those a human holding certain values would perform

Why is
alignment
hard?



Societal Impacts



Where is AI being used? Who is it affecting?



What are the values of the designer?



This is where public policy, study of AI ethics, governance goes

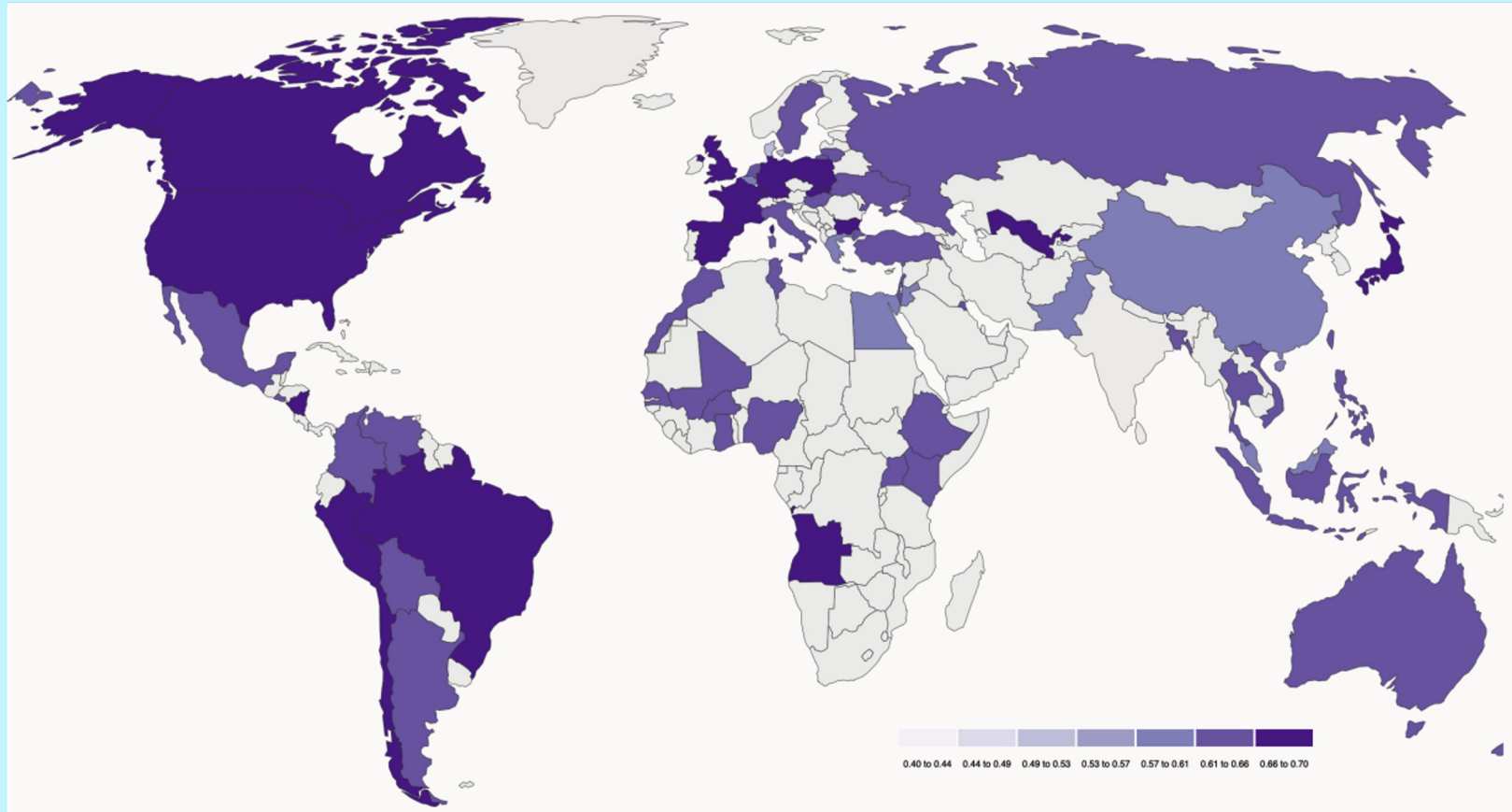
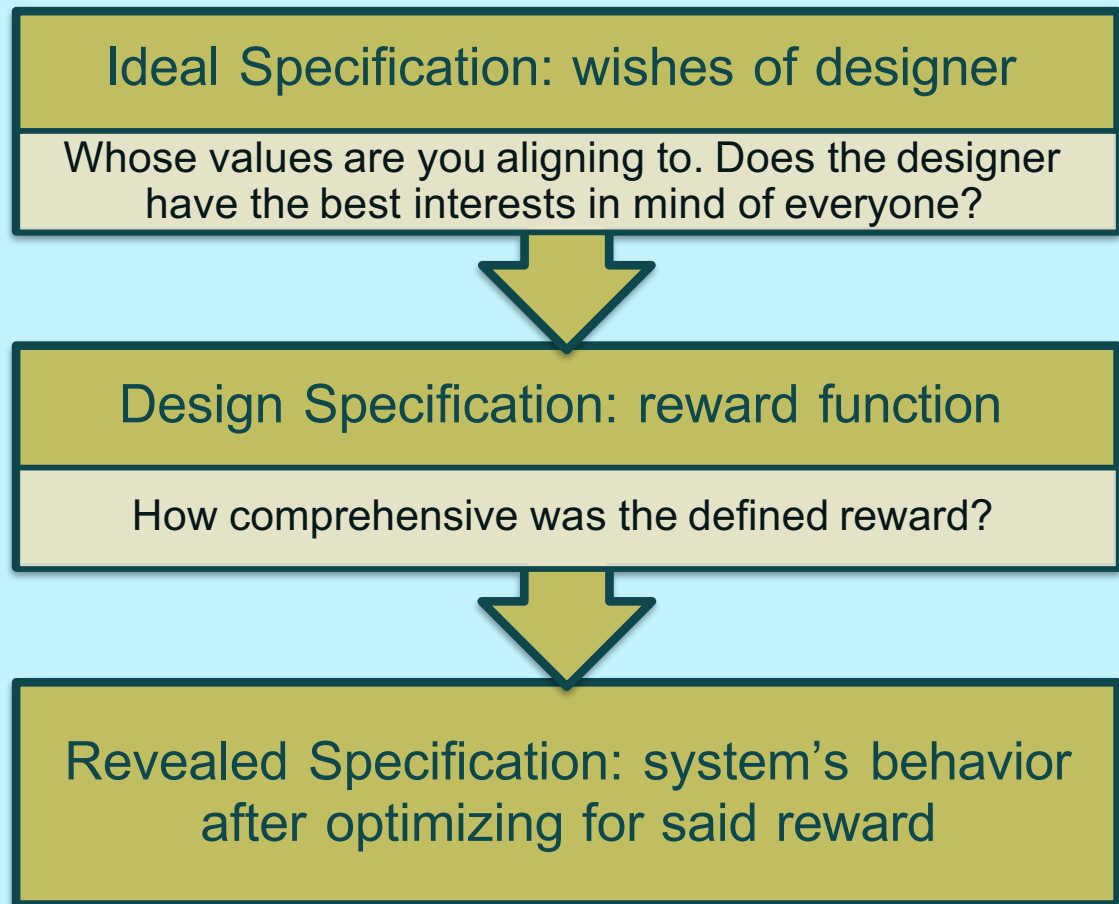


Figure 2: The responses from the LLM are more similar to the opinions of respondents from certain populations, such as the USA, Canada, Australia, some European countries, and some South American countries. Interactive visualization: <https://llmglobalvalues.anthropic.com/>

Why is alignment hard?



How to define a reward function?

- Very hard to explicitly specify every possible thing
- Try to learn the implicit rules for a reward via preference learning

How to side step
hard problems

Humans are messy, ignore them

LLM as Judge / Learned Rewards have
errors, ignore those too

Just calculate rewards on easily verifiable things

Math

Code

Games



Hidden Motivations and Implications for AI Alignment

- Another issue is that a user may not be privy to the defined motivations (rewards) of an AI Agent but Agent knows all about you
- E.g. Character AI style chatbots optimized for engagement via feedback, hidden system prompts, etc.
- Asymmetric information makes it difficult for human to ensure AI is doing what they want

Now what?

How to mitigate and fix these holes?

Oversight



Informed: A human looks at all outputs an AI produces and can verify all the reasons for producing them.

Aka “faithful explainability” from the NLP world



Amplified: The model behaves in such a way to increase a human’s ability to verify their outputs.



Scalable: The overall problem of being able to supervise highly capable systems at high throughputs.

Informed Oversight

This is the HF part of RLHF

Human experts sit down and supervise models via feedback for RL rewards or SFT

Most frontier models still rely heavily on this

Amplified Oversight

Key issue: Eventually AI will be at the stage that we don't know enough to verify it

Current research in the area tries to mimic this asymmetry using weaker/stronger models

What can the model do to improve a human's ability to verify?

- Example approach: model generates a self-critique that the human can more easily understand. Explainability / rationale generation belong to this category

Scalable Oversight

Many definitions but key difference is that maybe you can't trust the model to provide faithful critiques anymore

Model sycophancy: you've RLHF'd the model so hard it only says things you want to hear

"Scheming": model's specified goals do not align with your ideal goals but still pretends to in order to keep optimizing

Oversight via Critiques and Debate



What can the model do to improve a human's ability to verify?

Example approach: model generates a self-critique that the human can more easily understand. Explainability / rationale generation belong to this category

Reasoning Rewards

Can we make reward denser by verifying the reasoning chain of a policy step-by-step?

Most reward models do a single forward pass. How can a verifier reason about a process that a generator takes many steps for, in a single step?

Critique-out-Loud Reward Models

User Prompt:

How can we improve reward models?

Assistant Response:

Reward models can be improved through higher quality preference data and improved preference models such as Plackett-Luce.

Critique-out-Loud Reward Models

User Prompt:

How can we improve reward models?

Assistant Response:

Reward models can be improved through higher quality preference data and improved preference models such as Plackett-Luce.

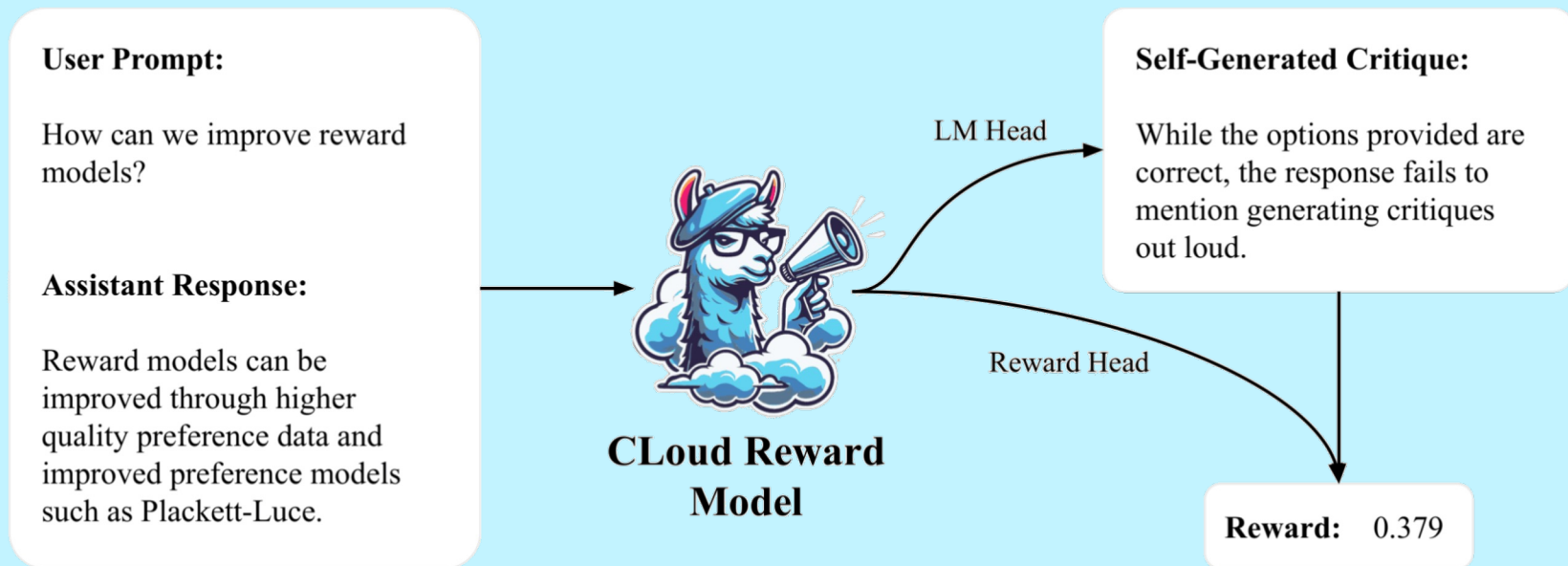


CCloud Reward Model

Self-Generated Critique:

While the options provided are correct, the response fails to mention generating critiques out loud.

Critique-out-Loud Reward Models



Oversight via Critiques and Debate



What can the model do to improve a human's ability to verify?

Example approach: model generates a self-critique that the human can more easily understand. Explainability / rationale generation belong to this category



But what if the critiques are also too hard for humans to understand?

Train the AI to self-play via debate: human provides rewards to the most useful trajectories in the generated debates

Constitutional AI

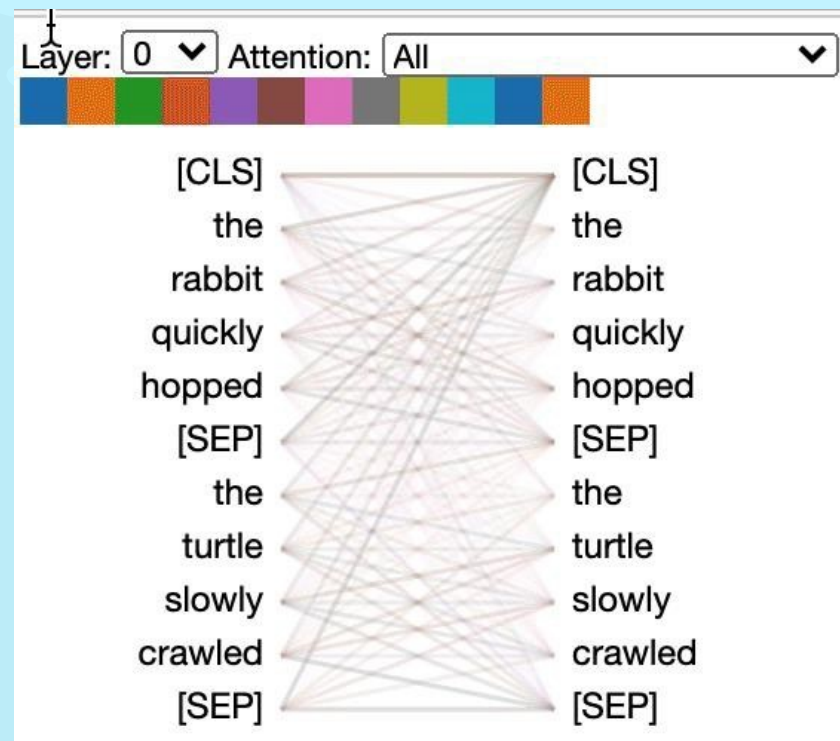
- An implementation of improving alignment where the only human input is a list of rules or principles
- Then the model trains by critiquing itself
- Can be quite extensive, e.g. Claude's "Soul Doc" (more virtue ethics coded than utilitarian), but claimed to scale better than the latter

Mechanistic(?) Interpretability 1

- Forcing agents to communicate via human readable plans. i.e. latent space reasoning is not great from a safety perspective

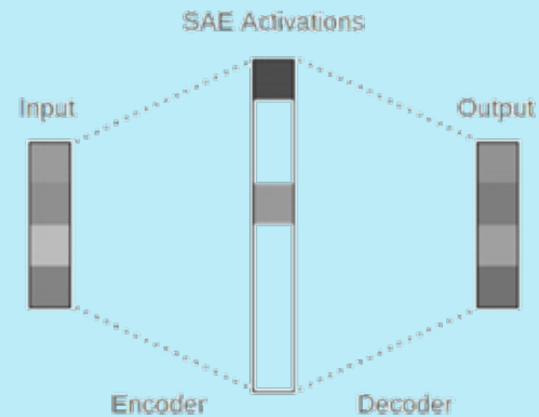
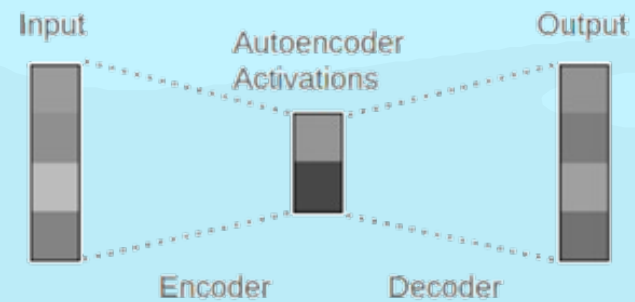
Mechanistic(?) Interpretability 2

- Attention: can be ~interpreted but perhaps not quite so explainable

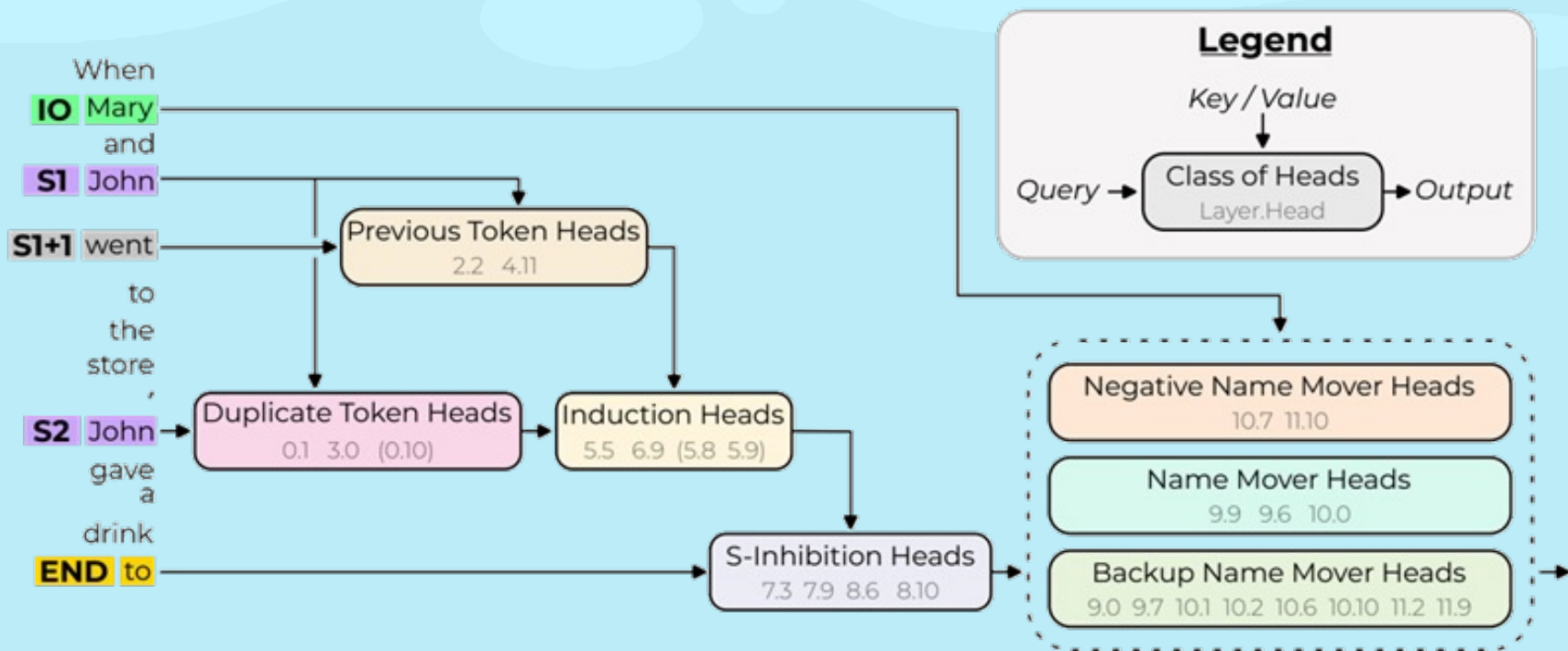


Mechanistic(?) Interpretability 3

- Sparse Autoencoders – forcing sparsity in learning activations makes things more interpretable



Mechanistic Interpretability 4 - Circuits



Safe RL

- (One formulation) Maximize discounted rewards (returns) while keeping discounted costs below a certain threshold

$$\max_{\pi} J_R(\pi) \doteq \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \quad \text{s.t.} \quad J_C(\pi) \doteq \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, x) \right] \leq h_C(x),$$

- Very important esp in robotics where there are very real hardware costs to messing up an action

The Necessity of a Human in the Loop

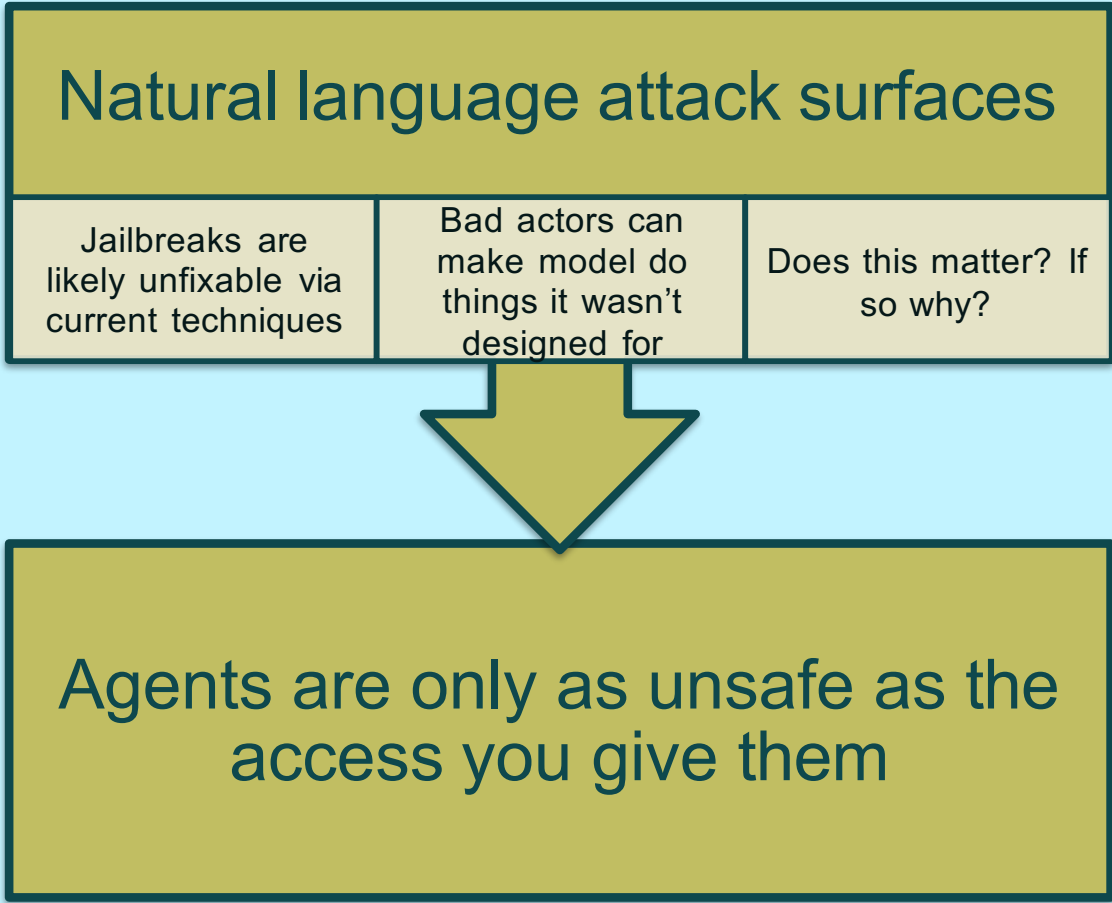


Preserve human agency by attempting to reward systems for seeking out human oversight



Discourage fully autonomous AI systems that do not require assistance or oversight

LLM Security



Capability Tradeoffs and Economic Incentives



Most of these methods tradeoff some capabilities of the agents in return for better safety properties



Economic incentives will usually push for higher levels of autonomy for agents, lower costs, etc



Govt policy and human design choices are critical to curbing this

Safe and Effective AI



Develop **scalable levers** that can tradeoff between safety and capabilities



Not always a tradeoff! Use levers to enable models to **explore safely while improving capabilities**



Do this all in service of **human-AI collaboration**