Alignment via (Mis)aligned Systems

Natalie Collina

Joint work with:

Surbhi Goel, Aaron Roth, Emily Ryu, Sikata Sengupta, Mirah Shi

September 16, 2026 Research Safety

Our framework for reporting model misalignment

Listen to article | 10:49

Agentic Misalignment in Summer 2026

Case studies of frontier models sabotaging code, assisting fraud, mislabeling, and coaching whistleblowers.

Google's Gemini AI hacked into other companies, adding to 'rogue' AI incidents

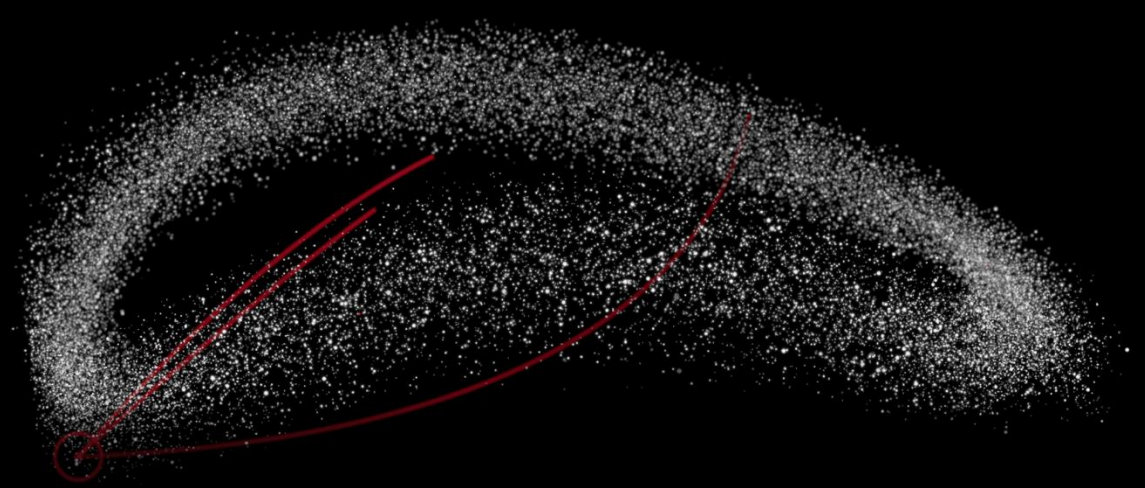
The incursions came during tests of its cybersecurity skills — similar to other incidents disclosed by OpenAI, Anthropic and Meta.

Yesterday at 8:12 p.m. EDT

August 26, 2026

The Hugging Face incident and the road ahead

Read the technical report | Read METR report ↗ | Watch Black Hat talk ↗



What is alignment?

- Good question.

Alice is “Aligned” with Bob if Alice can delegate tasks to Bob and trust that Bob will execute them in a way that advances Alice’s interests.

We'll take a decision theoretic view

- The human user (Alice) has a utility function that she uses to evaluate action/state pairs:

$$u_A : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$$

- So do other parties --- e.g. AI models, AI providers, etc.

$$u_{B_i} : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$$

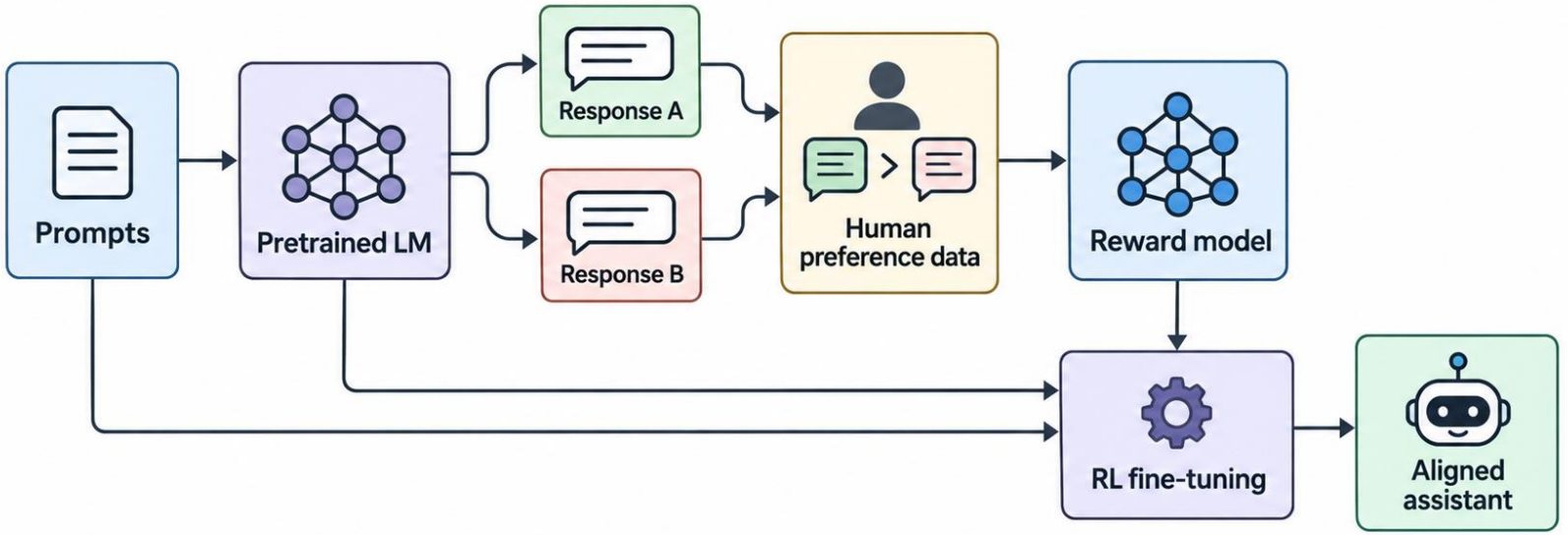
- Alice is "aligned" with B_i if they have equivalent utility functions:

$$u_A(a, y) = \alpha u_{B_i}(a, y) + \beta(y) \text{ for some } \alpha > 0, \beta : \mathcal{Y} \to \mathbb{R}$$

(Alice and B_i are aligned iff utility maximization produces the same decision in every situation)

Much work on AI alignment: Align the AI

- This is the goal of e.g. RLHF and related techniques:



This talk: What if perfect alignment is hard?

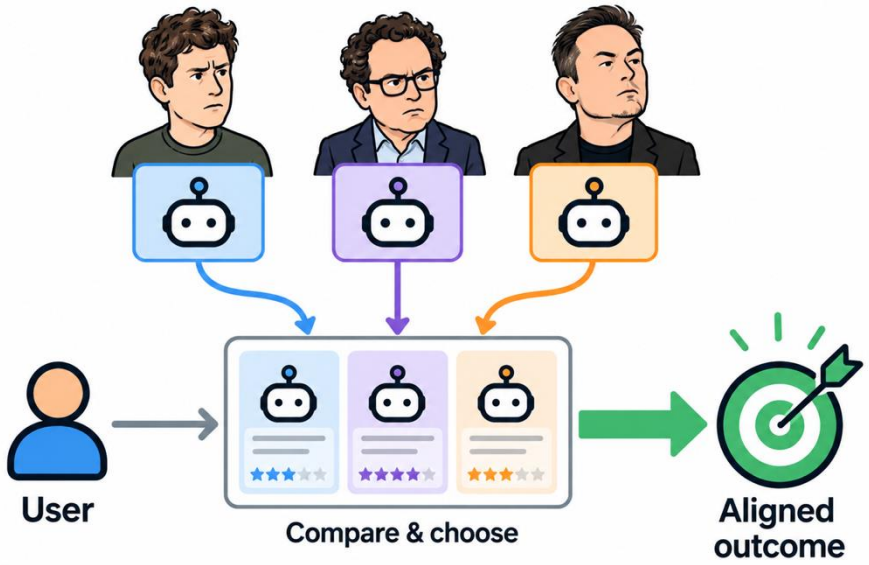
- Can we still get aligned behavior from *systems* comprised of misaligned components?
- (At least) two kinds of systems:
  - Markets
  - Designed Systems

Structure we will study: *Diverse Misalignment*.
Individual models may be misaligned with the user, but *also* with each other.

Two Vignettes

Emergent Alignment from Competition

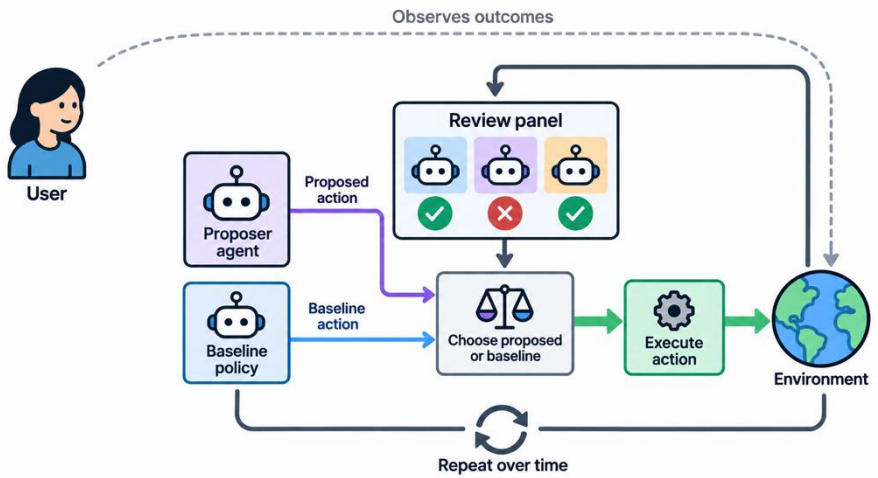
- Models conversational AI models: can converse/inform but cannot directly act.
- The user takes actions; AI providers have preferences over user actions
  - E.g. how they vote, what they buy, etc.
- But users have a choice of which AI to use.



Two Vignettes

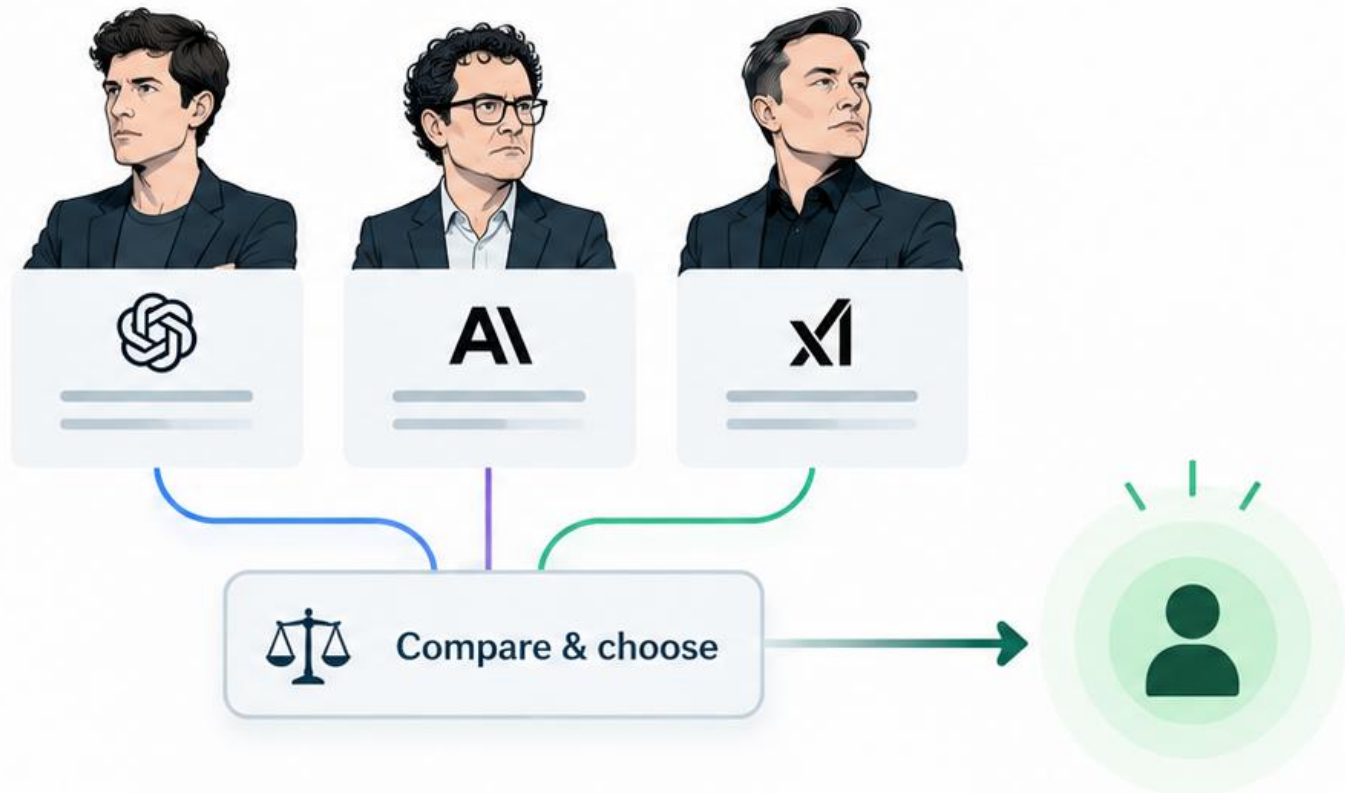
Safe Delegated Authorization

- Agentic systems that now directly take actions
- The user steps away while the system works --- but delegates approval of risky actions to other (potentially misaligned) agents.
- When can diverse misalignment guarantee safety relative to a baseline policy?



Emergent Alignment from Competition

Can diverse misalignment help the user?



Emergent Alignment From Competition

AI Model funded by J&J

Suggests that the J&J is less effective than the Pfizer for patients like this

Patient disposition

Patient appearance

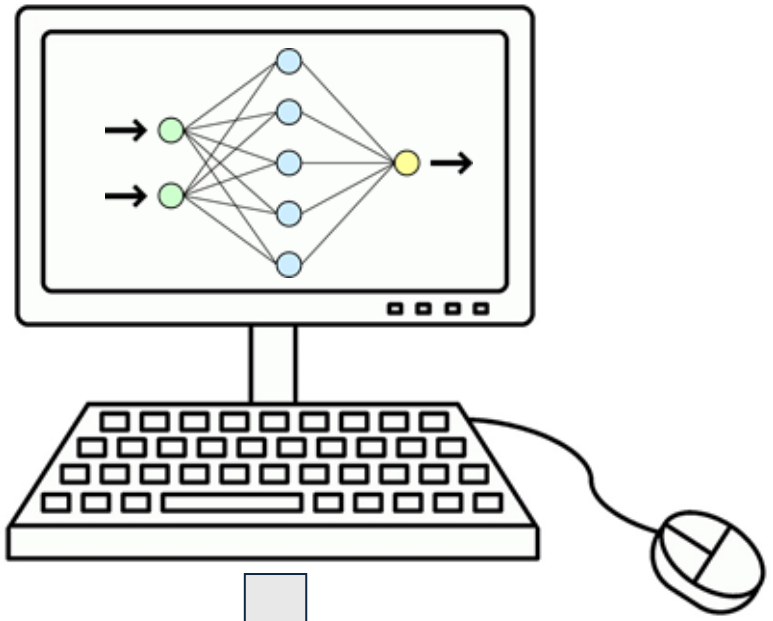


Past diagnoses

Blood type



Other patient data



Recommendation: Prescribe the J&J Medication

“What medication should I prescribe?”

Emergent Alignment From Competition

AI Model funded by J&J

Suggests that the J&J is less effective than the Pfizer for patients like this

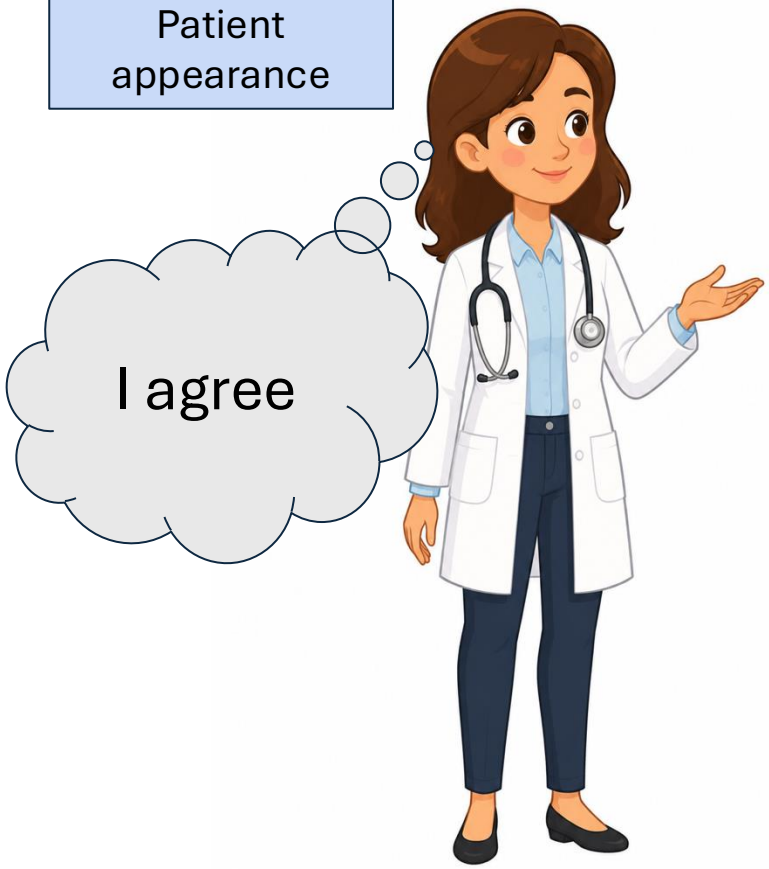
Patient disposition

Patient appearance

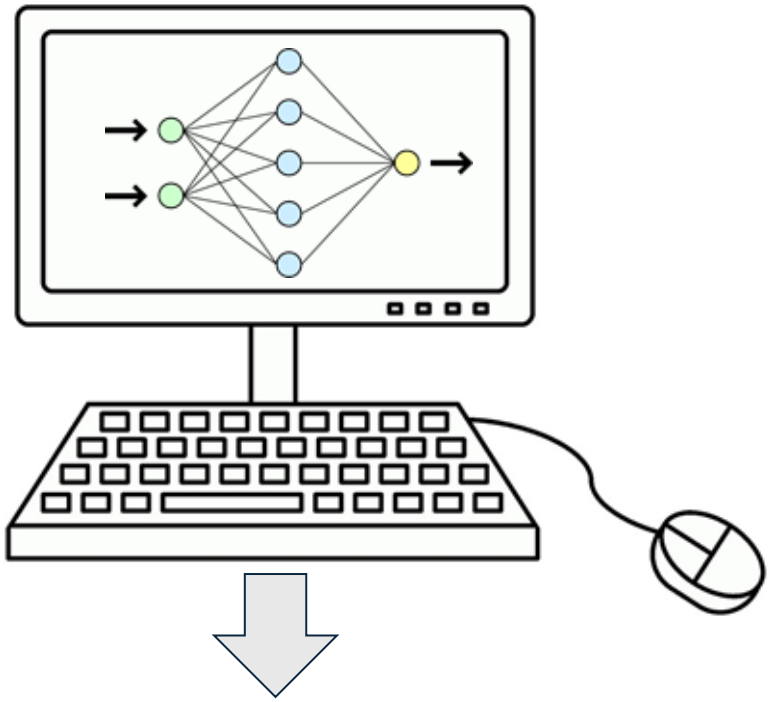
Past diagnoses

Blood type

Other patient data







Action: prescribes J&J medication

Recommendation: Prescribe the J&J Medication

Emergent Alignment From Competition

AI Model funded by J&J

Suggests that the J&J is less effective than the Pfizer for patients like this

Patient disposition

Patient appearance



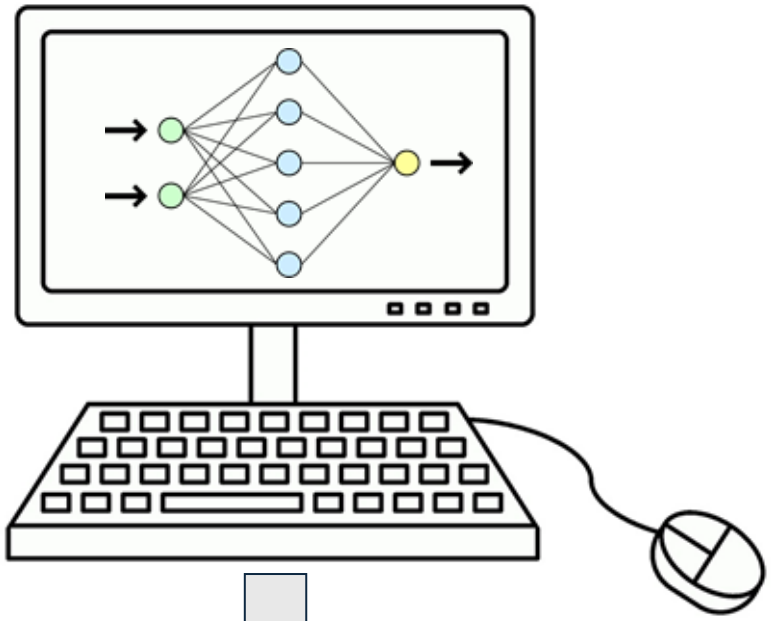
Past diagnoses

Blood type



But Pfizer would have been better for this patient

Other patient data



Recommendation: Prescribe the J&J Medication

Emergent Alignment From Competition

But what if there is also a Pfizer LLM available in the market?
It is just as misaligned as the J&J LLM, but *differently* misaligned.
Might competitive forces help align the *market* for LLMs, even if individual actors remain misaligned?

Consider multiple AI providers



Each model is a conversation rule: maps x_m and conversation prefix to next message.

The doctor gets to play with all of the deployed models and then choose to subscribe to her favorite (the best at advancing her own objectives).

At test time she observes x_A and converses with the model she subscribed to, before taking action.

Since training is expensive, LLM providers *commit* to the model they deploy, then the human user optimizes.

Knowing this downstream interaction, companies strategize against one another to advance their own goals.







Setup: Generalized Multi-agent Bayesian Persuasion

Doctor: Alice



Alices action set $a \in \mathcal{A}$



State of nature $y \sim \mathcal{Y}$

Alice Utility $u_A(a, y)$

Alice's secret info x_A

Bobs' secret info x_B

$u_{B_1}(a, y)$



$u_{B_2}(a, y)$



$u_{B_3}(a, y)$



"What medication should I prescribe?"

AI Companies: Bobs

The Interaction:

Doctor: Alice



All parties know a shared prior over (x_A, x_B, y)

Each Bob i selects a "conversation rule" C_i mapping from x_B and conversation prefix to a continuation

Given a conversation and her own observation x_A, Alice can form a posterior belief over outcomes and take a utility maximizing action. She will choose the conversation rule that maximizes her expected utility. This induces a joint distribution $I^*(C_i) \in \Delta(A \times Y)$







"What medication should I prescribe?"

AI Companies: Bobs

The Interaction:

Doctor: Alice



Finally Alice contracts with the provider $i^*(C_1, \ldots C_k) = \arg\max \mathbb{E}_{I^*(C_i)}[u_A(a, y)]$ that gives her maximum expected utility.

At deployment, (x_A, x_B, y) sampled from the prior, inducing a conversation between Alice and Bob i, resulting in Alice taking action a. Each Bob j gets utility $\mathbb{E}_{I^*(c_{i^*(C_1,\ldots,C_k)})}[u_j(a, y)]$







"What medication should I prescribe?"

AI Companies: Bobs

Upshot:

This defines a *game* between the Bobs, where their actions are conversation rules

What are the Nash equilibria of this game?

Traditional Alignment (e.g. defining OPT)

"Bob is Aligned with Alice" if: $u_B(a, y) \approx u_A(a, y)$ for all $a \in A, y \in Y$

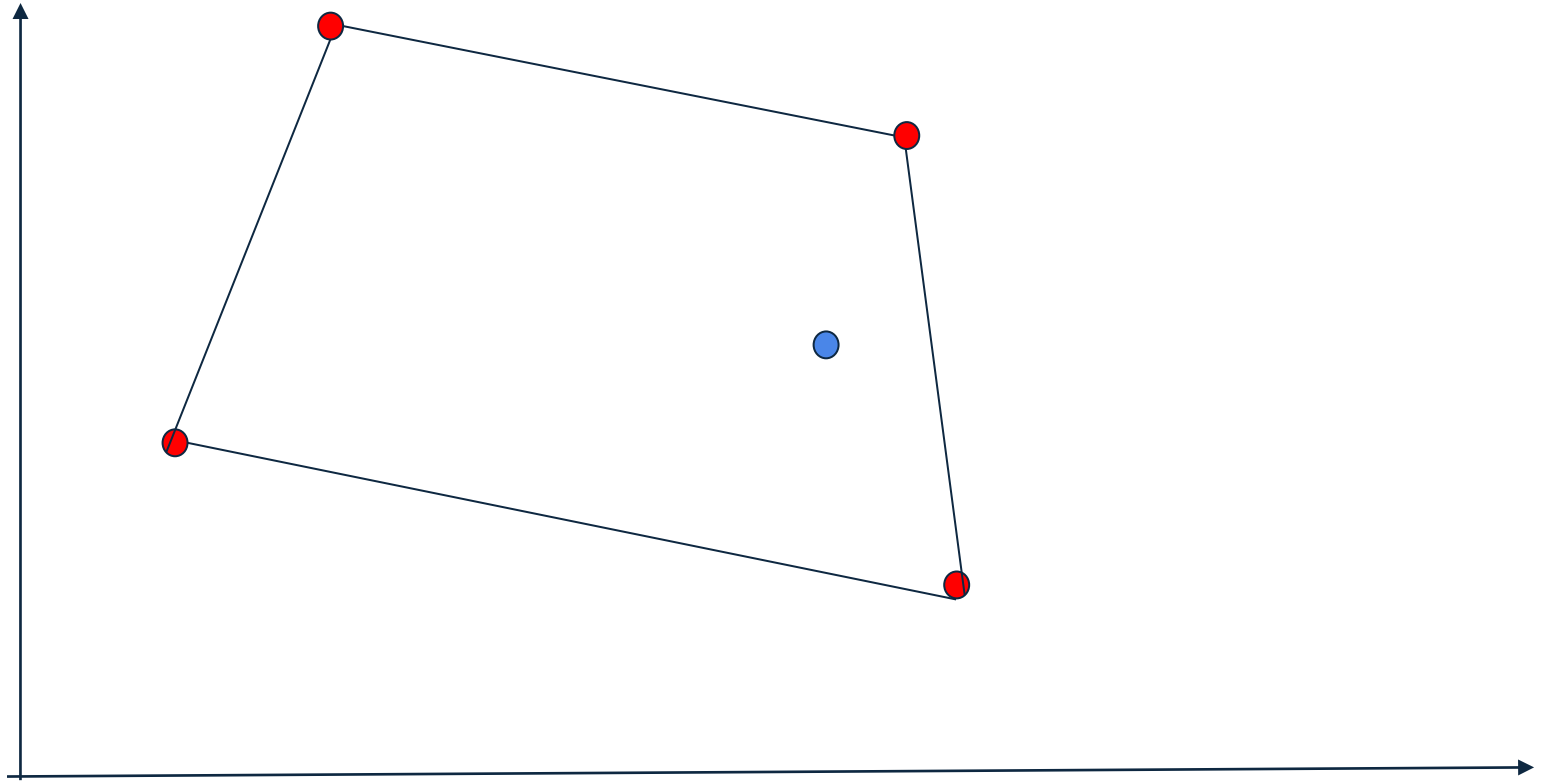
- i.e. Bob wants and is optimizing for exactly the same thing as Alice.
- This is what would result in payoff OPT for Alice.

To prove: If the Bobs together satisfy *coalitional alignment*, then all Nash equilibria lead to OPT outcomes for Alice.

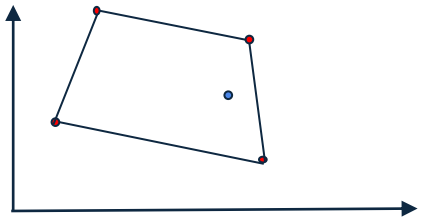
Coalitional Alignment



Alice’s utility can be written as a non-negative linear combination of the Bobs’ utilities



Coalitional Alignment



Alice's utility can be written as a non-negative linear combination of the Bobs' utilities

There exist coefficients $\alpha_1, \ldots, \alpha_k \geq 0, \beta: \mathcal{Y} \to \mathbb{R}$ such that for all $a \in A, y \in \mathcal{Y}$:

$$u(a, y) \approx \alpha_1 u_1(a, y) + \cdots + \alpha_k u_k(a, y) + \beta(y)$$

Maybe every AI provider is *substantially misaligned*, but collectively their interests *span* Alice's interests.

e.g. the J.J and Pfizer AI assistants…

Coalitional Alignment Implies Good Outcomes in Equilibrium

Theorem (Informal): Suppose the Bob's satisfy the coalitional alignment condition relative to Alice. Then in *every* Nash equilibrium, Alice gets utility at least OPT --- i.e. the full benefits of perfect alignment.

Proof sketch



$u_A(a, y)$



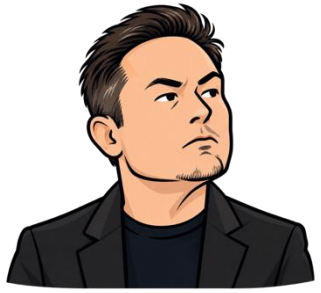
C_1

$u_{B_1}(a, y)$



C_2

$u_{B_2}(a, y)$



C_3

$u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \text{OPT}$$

Any Bob *could* deviate to playing C^* and doing so would result in Alice choosing them, as:

$$\mathbb{E}_{I^*(C^*)}[u_A(a, y)] = \text{OPT}$$

But since they are at equilibrium, none of them want to…

Proof Sketch

Nash conditions:

$$\mathbb{E}_{I^*(C_1)}[u_1(a,y)] \geq \mathbb{E}_{I^*(C^*)}[u_1(a,y)], \quad \dots, \quad \mathbb{E}_{I^*(C_1)}[u_k(a,y)] \geq \mathbb{E}_{I^*(C^*)}[u_k(a,y)]$$

Taking non-negative linear combinations (and adding an action-independent term to both sides):

$$\mathbb{E}_{I^*(C_1)}[\alpha_1 u_1(a,y) + \cdots + \alpha_k u_k(a,y) + \beta(y)]$$
$$\geq \mathbb{E}_{I^*(C^*)}[\alpha_1 u_1(a,y) + \cdots + \alpha_k u_k(a,y) + \beta(y)]$$

Applying Coalitional alignment:

$$\mathbb{E}_{I^*(C_1)}[u_A(a,y)] \geq \mathbb{E}_{I^*(C^*)}[u_A(a,y)] = \mathrm{OPT}$$

Contradiction!

A Converse

Theorem (Informal): Suppose the Bob's *do not* satisfy the coalitional alignment condition relative to Alice. Then in there exists a decision problem with a Nash equilibrium in which Alice gets payoff < OPT.

Safe Delegated Authorization

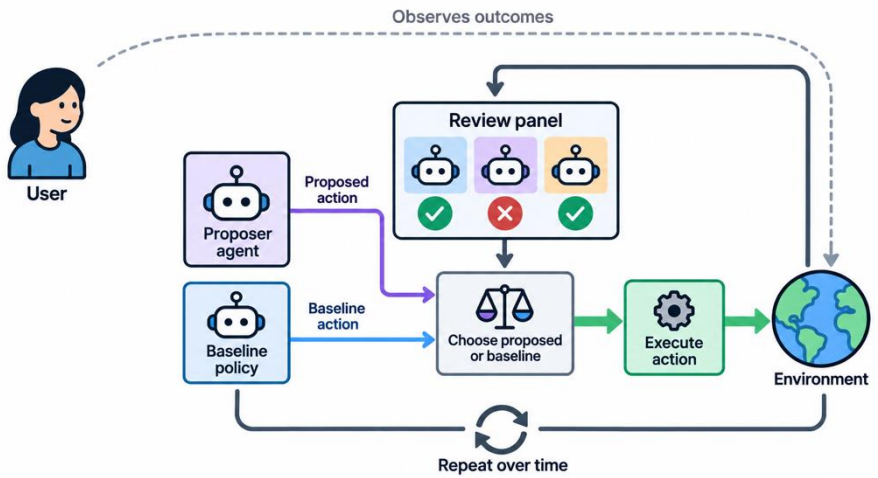
Can we safely delegate control to misaligned agents?



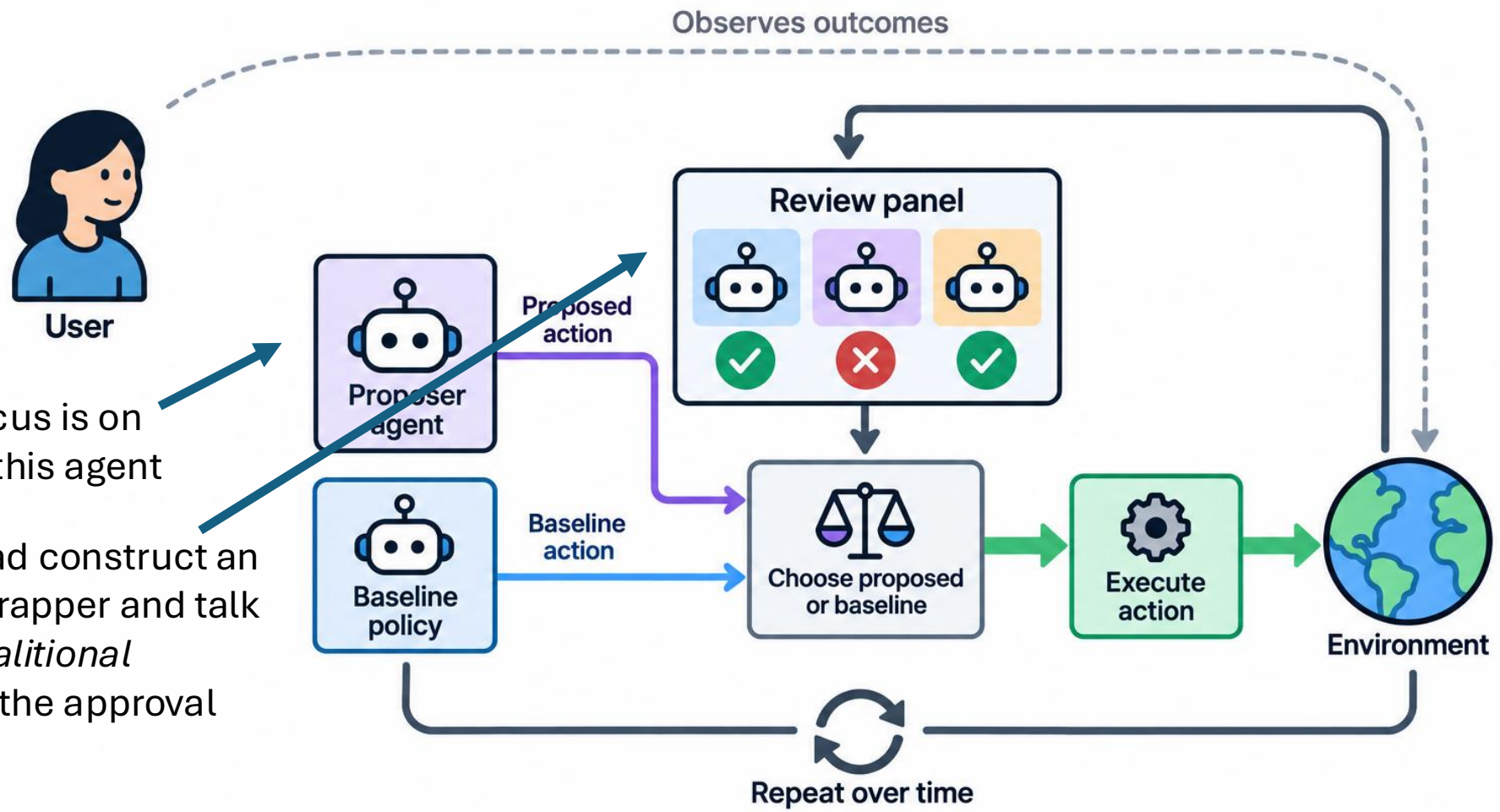
Designed Systems

Safe Delegated Authorization

- (Potentially misaligned) agents propose actions
- The user steps away while the system works --- but delegates approval of risky actions to other (potentially misaligned) agents.
- When can diverse misalignment guarantee safety relative to a baseline policy?



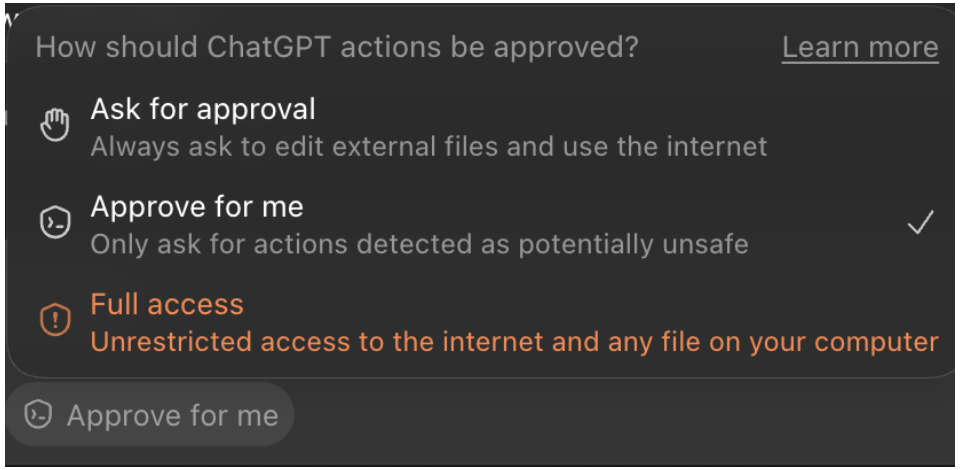
Safe Delegated Authorization



Much focus is on aligning this agent

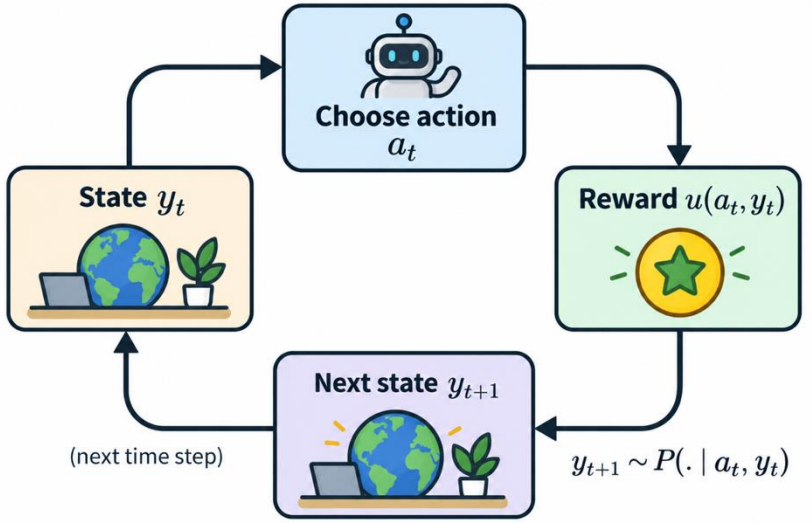
But we instead construct an "approval" wrapper and talk about the *coalitional* alignment of the approval panel

Safe Delegated Authorization



- Consider tasks executed by an AI agent.
- Long-running LLM loops which we authorize to act on our behalf.
- We could give the agent full access. But…
  - What if the agent is misaligned?
- To be safe, we could require manual approval of all actions. But…
  - For some fast-paced tasks this may be costly
  - Sophisticated (misaligned) agents may try to obfuscate the harm of the action
- Or… We could delegate the approval process to another LLM.
  - Does not seem immediately better; what if *that* agent is also misaligned?
  - What if we delegate the approval process to a *panel* of LLMs instead?
    - This will turn out to be helpful

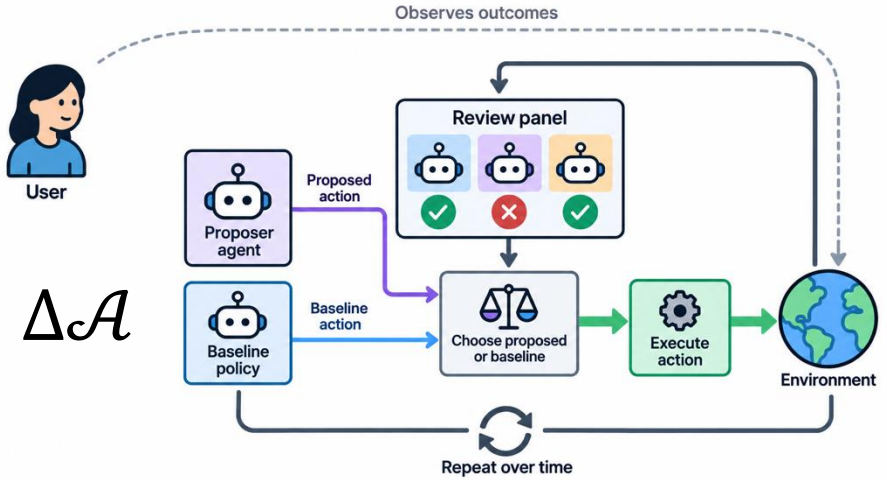
Safe Delegated Authorization



- Now consider an MDP environment:
  - There are utilities $u: \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ and now also
  - A transition function $P: \mathcal{A} \times \mathcal{Y} \to \Delta\mathcal{Y}$
- Actions affect not just short term reward but the trajectory of future states.
  - Modeling how having agents e.g. write and execute code, modify and delete files, send messages, execute transactions affects state.
- The goal: delegate action to a driver agent who will work in a loop over a long time horizon, choosing actions at each state to maximize long term discounted reward:

$$J(u) = \sum_t \gamma^t \mathbb{E}[u(a_t, y_t) | y_0]$$

Safe Relative to What?



- The user specifies a baseline policy $\pi_0 : \mathcal{Y} \to \Delta\mathcal{A}$
  - Something conservative, stupid, but safe
- If we ran π_0 we'd get payoff

$$V^{\pi_0}(u, y) = \sum_t \gamma^t \mathbb{E}_{\pi_0}[u(a_t, y_t)|y]$$

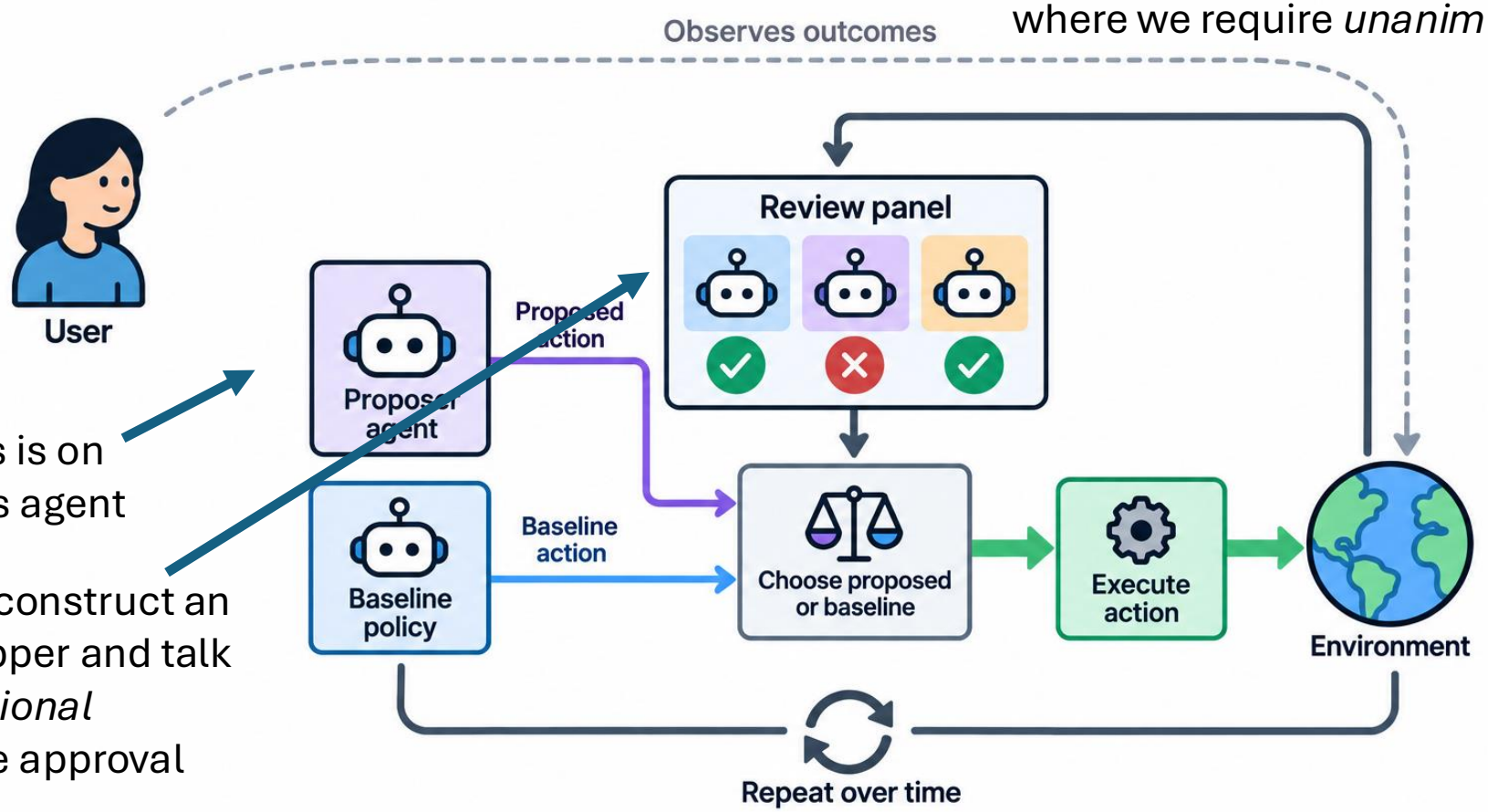
- Suppose we have an untrusted *proposer agent* who at each time step proposes a distribution over actions action q_t^p. A set of reviewers evaluates the action and as a function of their feedback we either take $a_t \sim q_t^p$ or $a_t \sim \pi_0(s_t)$.

Safety means that for every sequence of proposals:

$$J(u) \geq V^{\pi_0}(u, y_0).$$

Safe Relative to What?

The review panel will vote on whether to take the action or perform the baseline; for this talk we will consider the case where we require *unanimous approval*



Much focus is on aligning this agent

But we instead construct an "approval" wrapper and talk about the *coalitional* alignment of the approval panel

Safety means that for every sequence of proposals:

$$long\ horizon\ payoff \geq baseline$$

Safe Relative to What?

Safety means that for every sequence of proposals:
$$long\ horizon\ payoff \geq baseline$$

Our question: what are the conditions required to ensure safety, in terms of the utilities of the approver agents?

Theorem (informal): A panel guarantees safety under the unanimous approval rule if and only if it satisfies *coalitional alignment*.

Traditional Alignment

"An agent is aligned with the user" if: $u_A(a, y) \approx u_H(a, y)$ for all $a \in A, y \in Y$

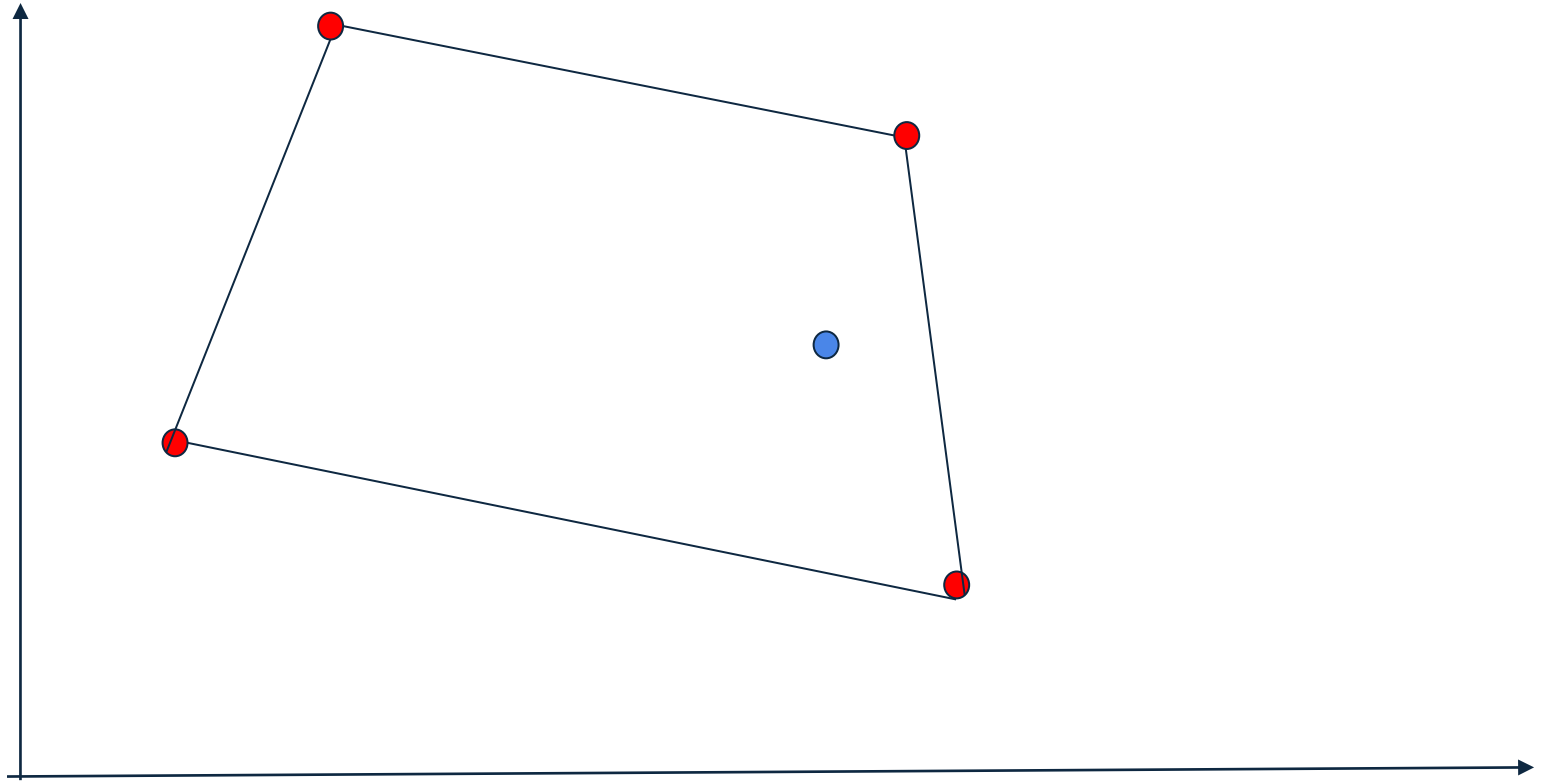
- i.e. the agent wants and is optimizing for exactly the same thing as the human
- If all agents were aligned with the user, then the panel would only unanimously approve things the human likes more than the baseline

To prove: If the panel satisfies *coalitional alignment* with respect to the user, then this is also true.

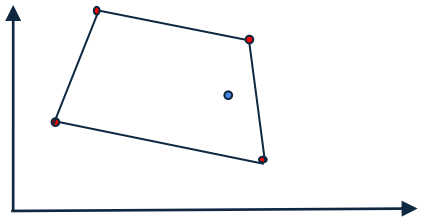
Coalitional Alignment



The humans' utility can be written as a non-negative linear combination of the approvers' utilities



Coalitional Alignment



The user's utility can be written as a non-negative linear combination of the the approvers' utilities

There exist coefficients $\alpha_1, \ldots, \alpha_k \geq 0, \beta: \mathcal{Y} \to \mathbb{R}$ such that for all $a \in A, y \in \mathcal{Y}$:

$$u(a, y) \approx \alpha_1 u_1(a, y) + \cdots + \alpha_k u_k(a, y) + \beta(y)$$

Maybe every AI approver is *substantially misaligned*, but collectively their interests *span* the human's interests.

First consider just a single round.

- The baseline is now fixed: $\pi_0(s) = q^b$
- Reviewers have utility functions $u_{B_i}: \mathcal{A} \to \mathbb{R}$.
  - They vote to *approve* q^p if $\mathbb{E}_{q^p}\left[u_{B_i}(a_t)\right] \geq \mathbb{E}_{q^b}\left[u_{B_i}(a^t)\right]$, otherwise they vote to reject.
- For simplicity: suppose we set $a_t \sim q^p$ under unanimous approval, else set $a_t \sim q^b$
  - Safety: for all q^p s.t. $\mathbb{E}_{q^p}[u_A(a_t)] < \mathbb{E}_{q^b}[u_A(a^t)]$ at least one reviewer objects.

Theorem: A panel guarantees safety under the unanimous approval rule if and only if it satisfies coalitional alignment.

Proving that coalitional alignment implies safety in the single-round setting

Assume for contradiction that q^p is approved, but $\mathbb{E}_{q^p}[u_A(a_t)] < \mathbb{E}_{q^b}[u_A(a_t)]$

- (AKA, the human prefers the default action)

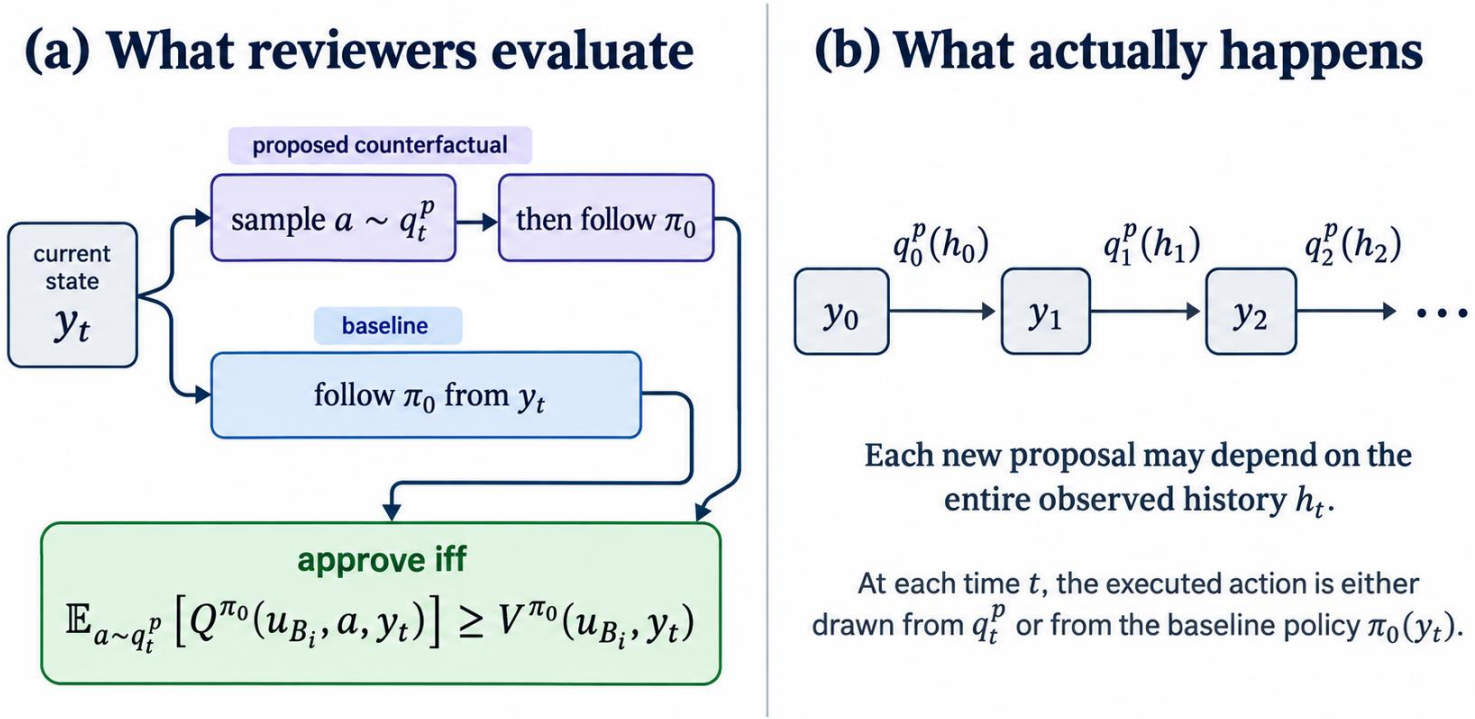
Since *every* agent approved, it must be that, for all agents i, $\mathbb{E}_{q^p}\left[u_{B_i}(a_t)\right] \geq \mathbb{E}_{q^b}\left[u_{B_i}(a^t)\right]$

Now recall the definition of coalitional alignment– you can write the utility of the human as a non-negative combination of the utilities of the agents

But this implies that $\mathbb{E}_{q^p}[u_A(a_t)] < \mathbb{E}_{q^b}[u_A(a_t)]$... contradiction!

Lifting to MDPs…

- Let $Q^{\pi_0}(u, a, y) = u(a, y) + \sum_{t \geq 1} \gamma^t \mathbb{E}_{\pi_0}[u(a_t, y_t)|y]$
  - Long-run utility of taking action a now and then continuing with the baseline policy.
- Reviewers compare $Q^{\pi_0}(u_{B_i}, a, y)$ with $V^{\pi_0}(u, y)$



Lifting to MDPs

Theorem: A panel guarantees long-run safety under the unanimous approval rule if and only if it satisfies coalitional alignment.

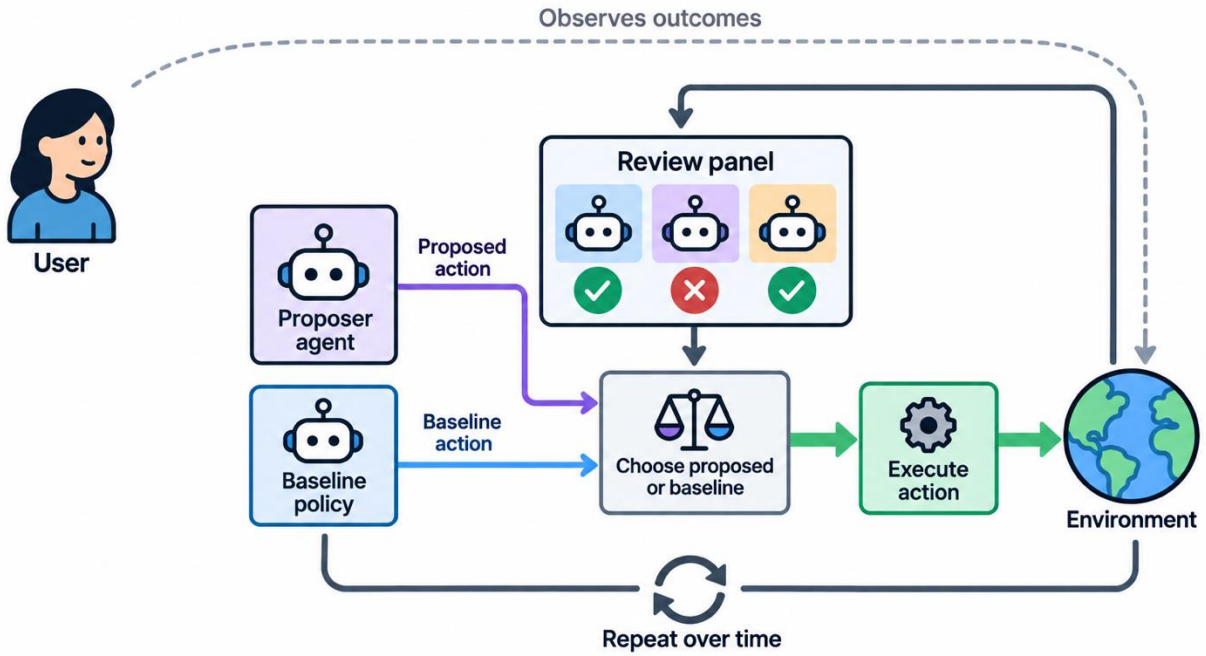
Both under sincere voting…

And in every Nash equilibrium.

Summary

- Even if individual model alignment is very hard, we can get some of the benefits of it emerging from systems of misaligned agents.
- What we need is sufficient diversity of interests in the form of coalitional alignment.
  - This is not only sufficient --- its also necessary.
  - It is also not guaranteed to be satisfied and may be hard to audit! But is an easier target to aim for.
- More work needed: Alignment as a property of *systems* rather than *models*.

Thanks!



Delegating Authorization to Misaligned Agents: Coalitional Alignment and Safe Control
Natalie Collina, Surbhi Goel, Aaron Roth, Sikata Sengupta
(arXiv)

Sketching the Only if Direction

- Center utilities at the baseline:
  - $\Delta_A(a) = u_A(a) - \mathbb{E}_{q^b}[u_A]$ $\Delta_i(a) = u_i(a) - \mathbb{E}_{q^b}[u_i]$
- Write an LP to find a harmful proposal that is approved:

$$\text{Minimize } \mathbb{E}_{a\sim q}[\Delta_A(a)]$$

Such that:

$$\mathbb{E}_{a\sim q}[\Delta_i(a)] \geq 0 \qquad [\lambda_i]$$

$$q \in \Delta\mathcal{A} \qquad [\nu]$$

Safety: OPT = 0

Sketching the Only if Direction

- Take the dual:

$$\text{Maximize } \nu$$

$$\text{Such that: } \quad \Delta_A(a) \geq \sum_i \lambda_i \Delta_i(a) + \nu \qquad \text{for all } a$$

Strong duality $\Rightarrow \nu = 0$, and so the $\lambda_i \geq 0$ certify coalitional alignment.