

AGI Safety Research Agendas

Rohin Shah, Center for Human-Compatible AI, UC Berkeley

What are people doing about AGI safety?



What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)

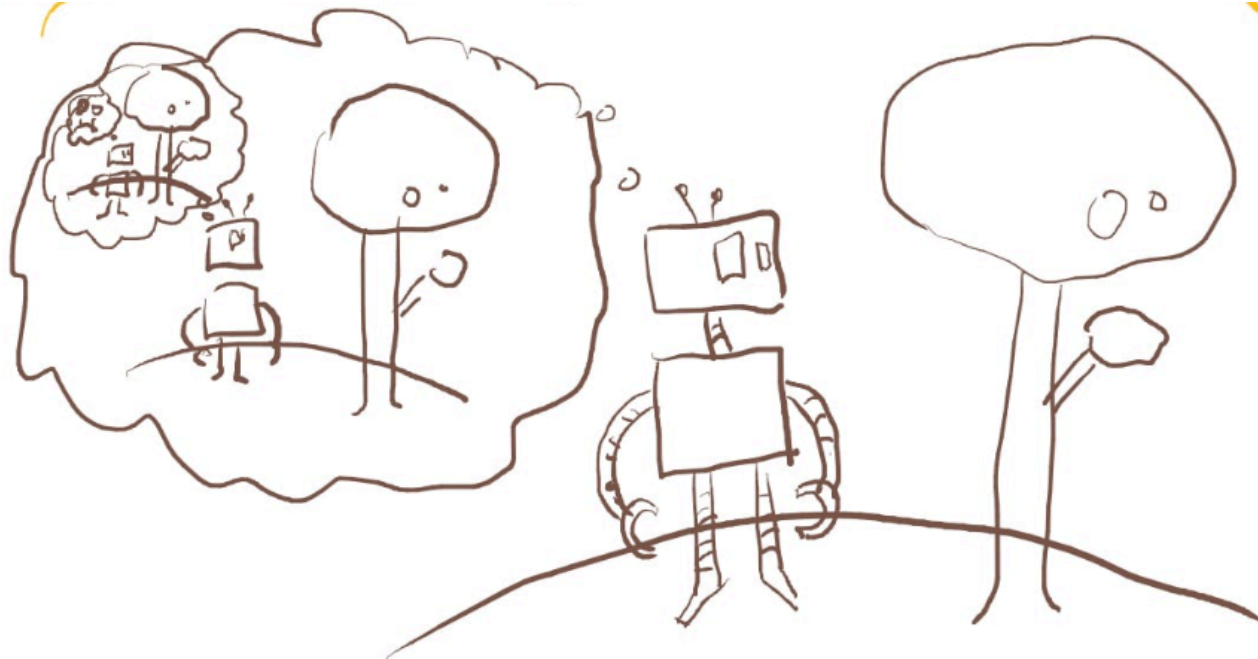
Embedded Agency

Decision Theory

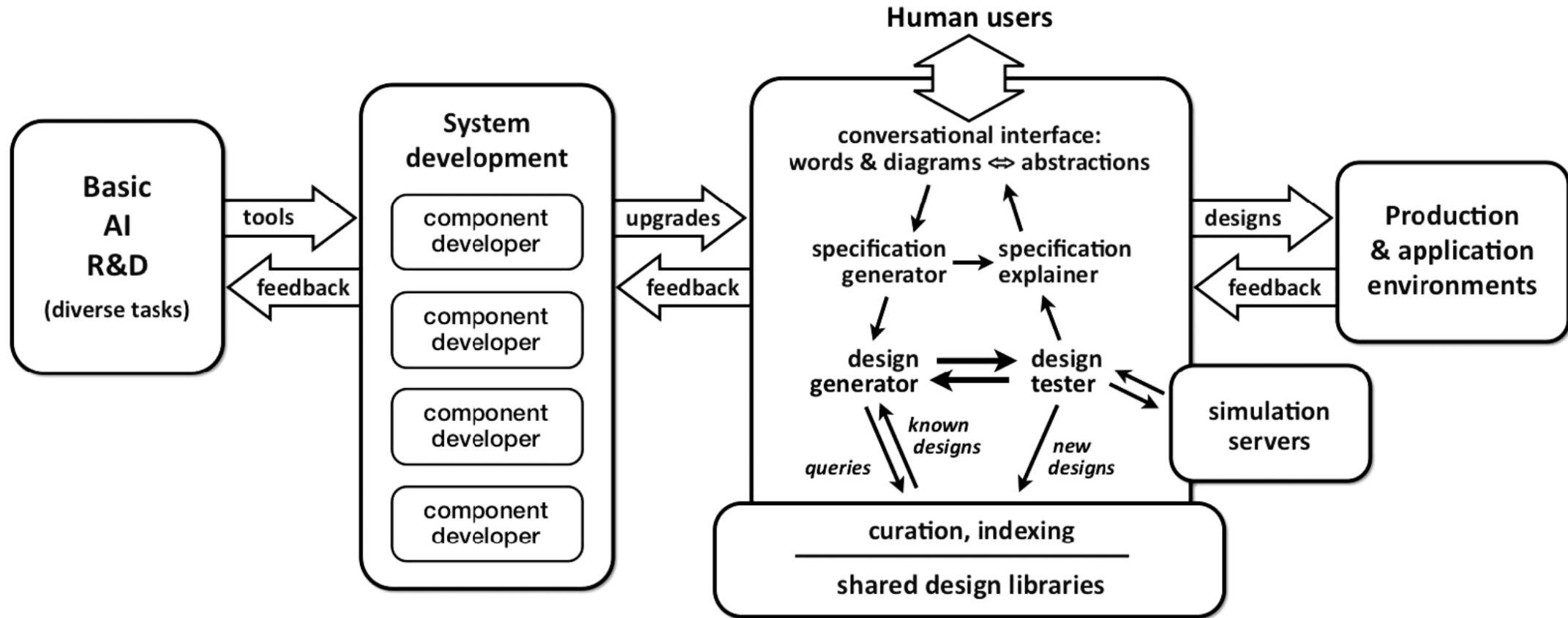
Embedded World-Models

Robust Delegation

Subsystem Alignment



Comprehensive AI Services



Human preferences are complex

Most behaviors are not catastrophic.

Most behaviors are not good.

So, good outcomes need a lot of information about humans, but avoiding catastrophic outcomes may not need much information.

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)

Impact measures



Vanilla



AUP



Tabular AUP



Relative Reachability



Starting State



Inaction



Decrease

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI
 - Having the right goal (Ambitious value learning)

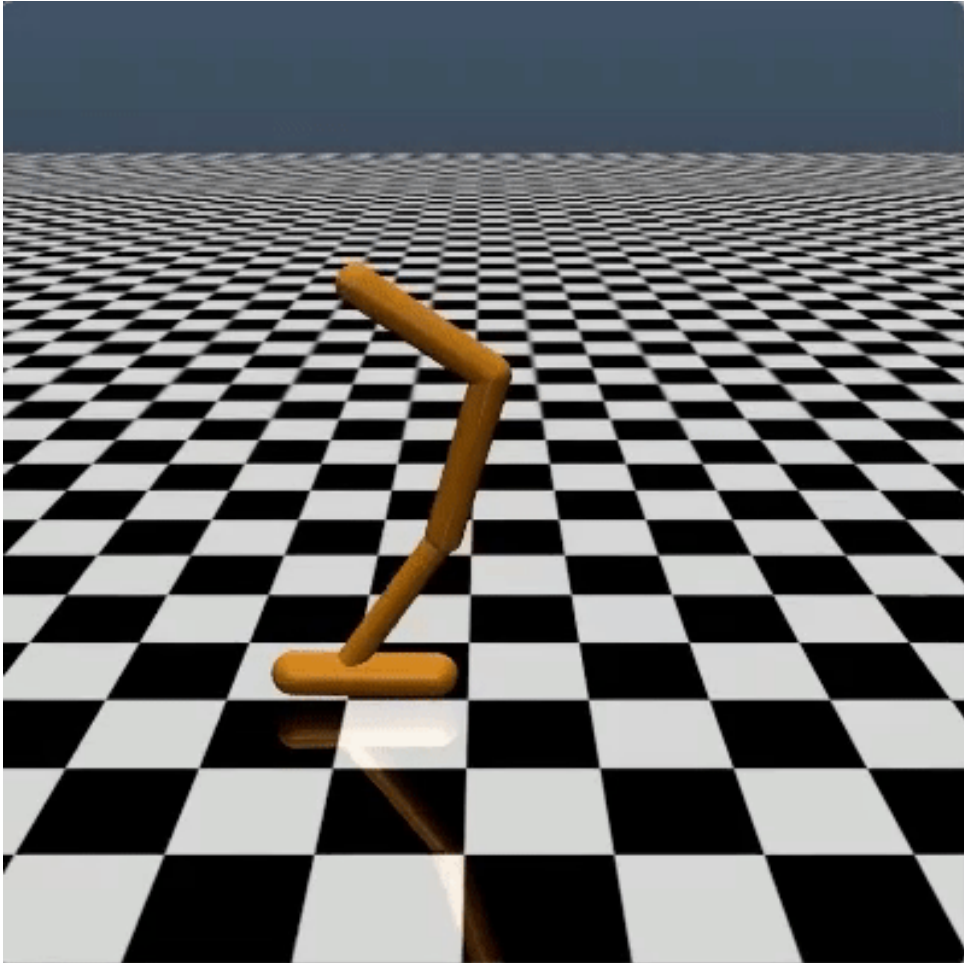
Ambitious value learning

- Infer values that can safely be loaded in superintelligent AI
- Challenge: How to deal with human biases?
- Make assumptions!
 - Analyze the human's decision-making algorithm
 - Notice facial expressions of regret

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI
 - Having the right goal (Ambitious value learning)
 - Human selection of behavior (Preference learning, IRL, Reward modeling)

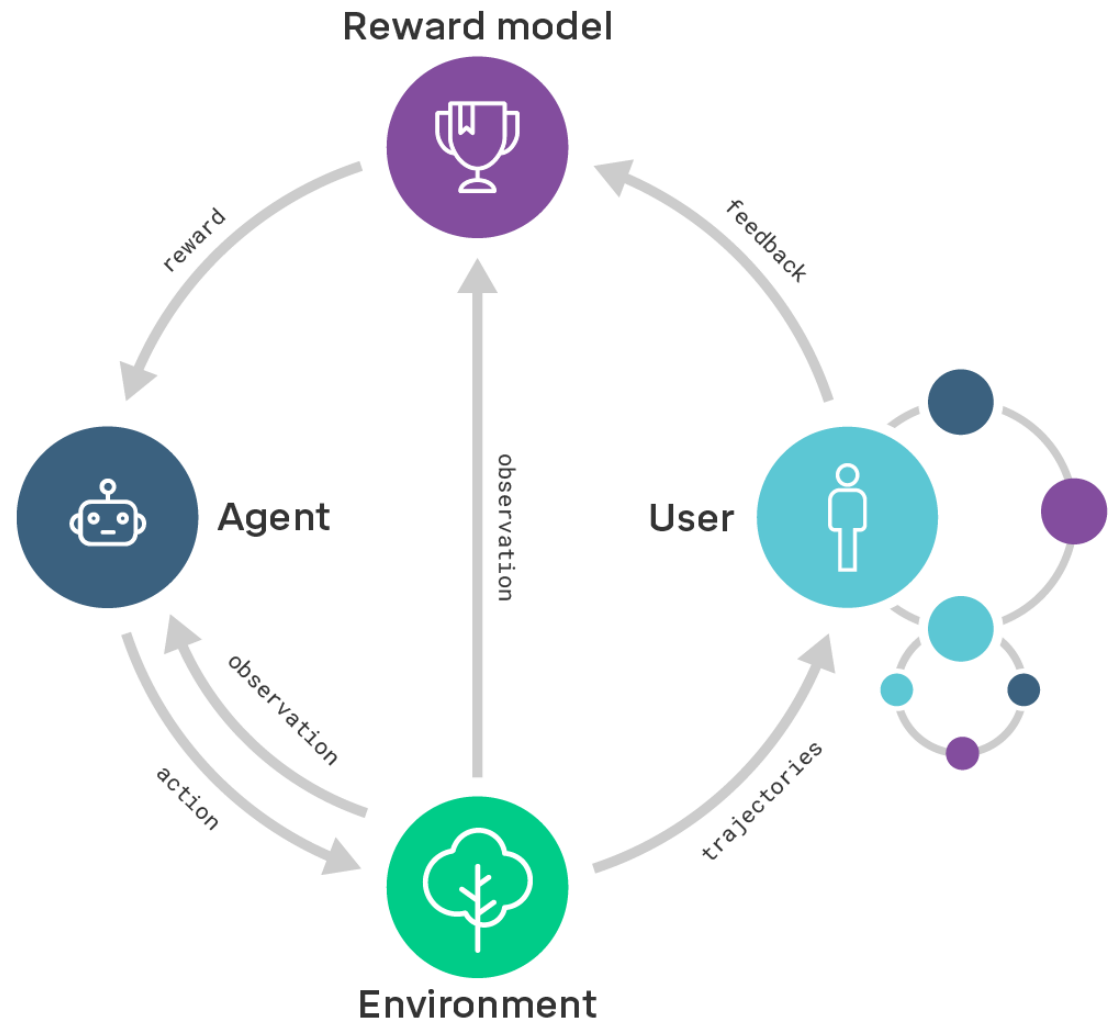
Preference learning



- Demonstrations
- Comparisons
- Ratings
- Stated reward function
- Initial state

Recursive reward modeling

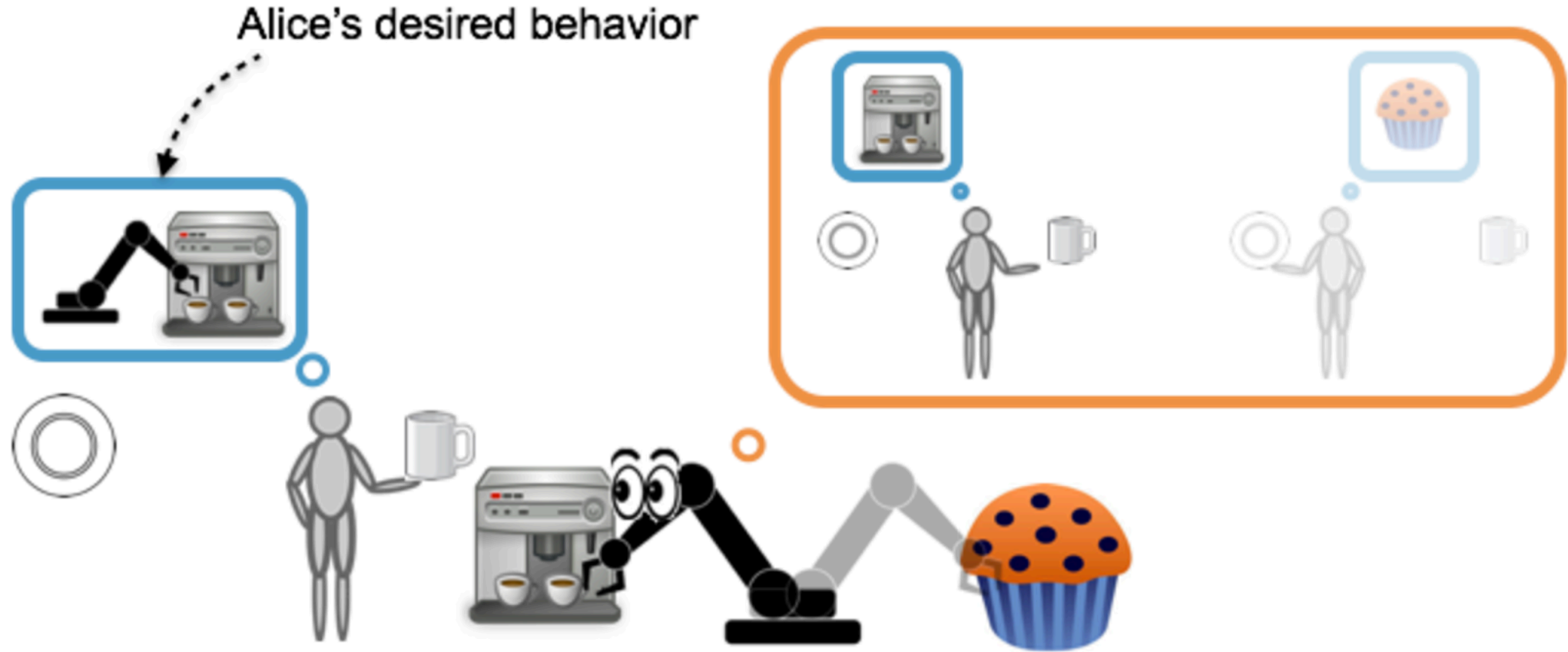
Scale to tasks that are hard to evaluate



What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI
 - Having the right goal (Ambitious value learning)
 - Human selection of behavior (IRL, Preference learning, Reward modeling)
 - Optimizing for *our* goals (Cooperative IRL)

Cooperative Inverse Reinforcement Learning



Rob observes Alice's actions to infer (and pursue) her desired goal.

What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI
 - Having the right goal (Ambitious value learning)
 - Human selection of behavior (IRL, Preference learning, Reward modeling)
 - Optimizing for *our* goals (Cooperative IRL)
 - Corrigibility (Iterated amplification, Debate, Factored cognition)

Corrigibility

How to make a beneficial AI system

```
graph TD; A[How to make a beneficial AI system] --> B[Definition]; A --> C[Optimization];
```

Definition

What behavior do we want?

Ambitious value learning

Optimization

How do we get that behavior?

Deep reinforcement learning

Corrigibility

How to make a beneficial AI system

```
graph TD; A[How to make a beneficial AI system] --> B[Motivation]; A --> C[Competence];
```

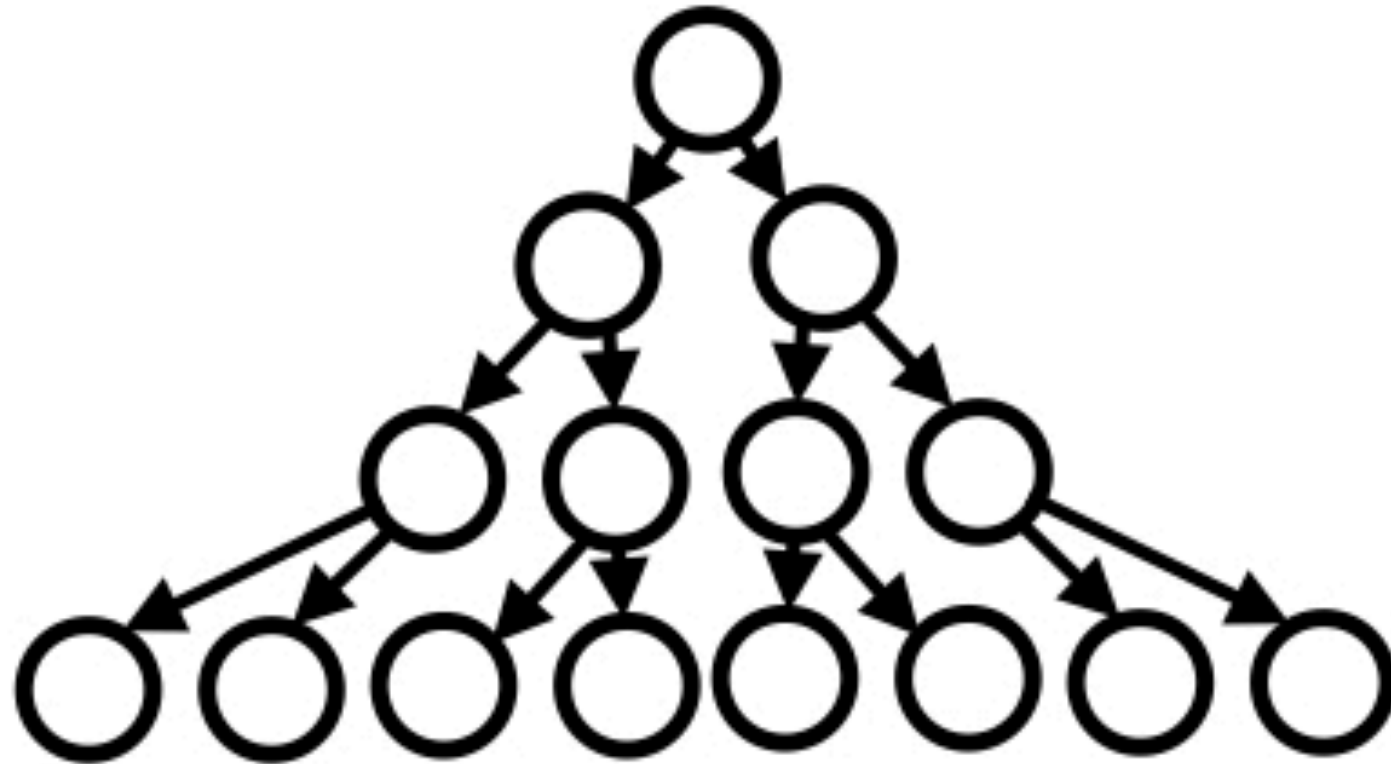
Motivation

Is our AI trying to help?

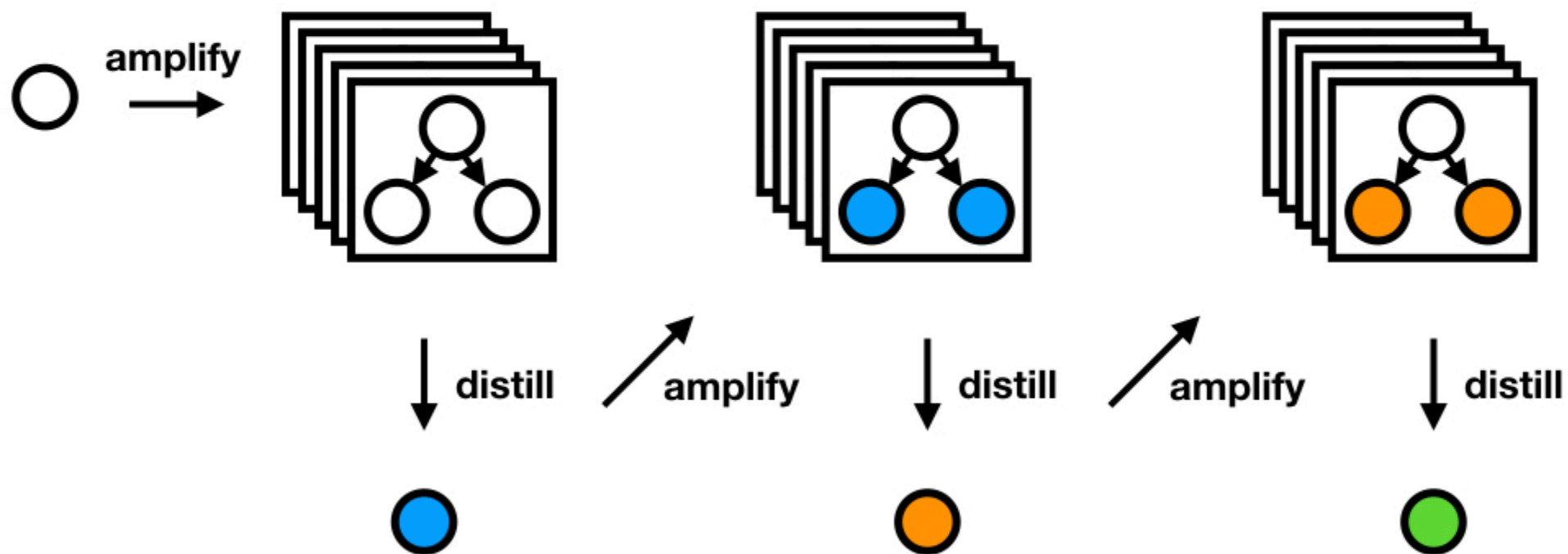
Competence

Is our AI good at helping?

Factored Cognition: Deliberation trees

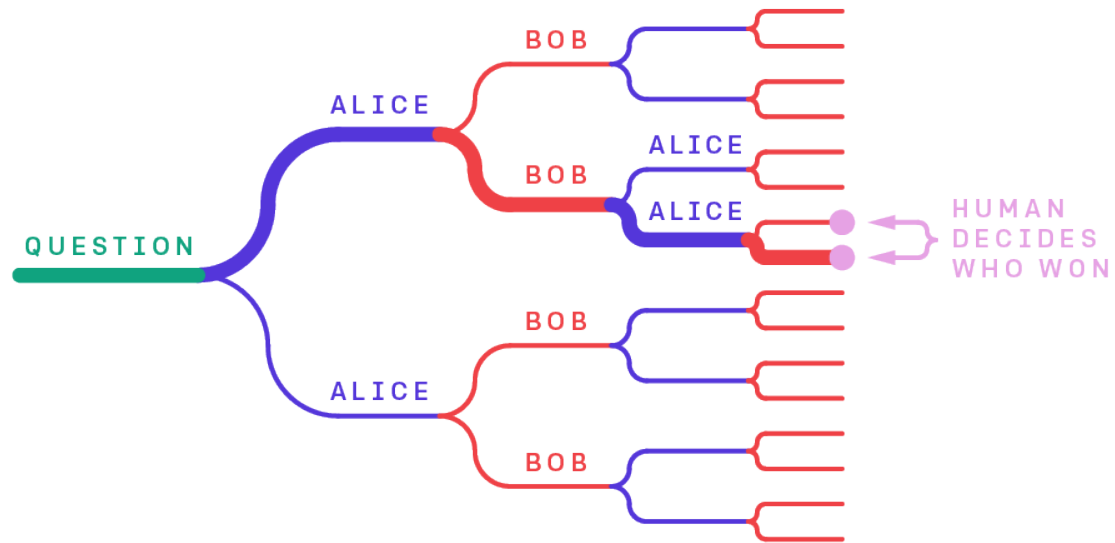


Iterated amplification

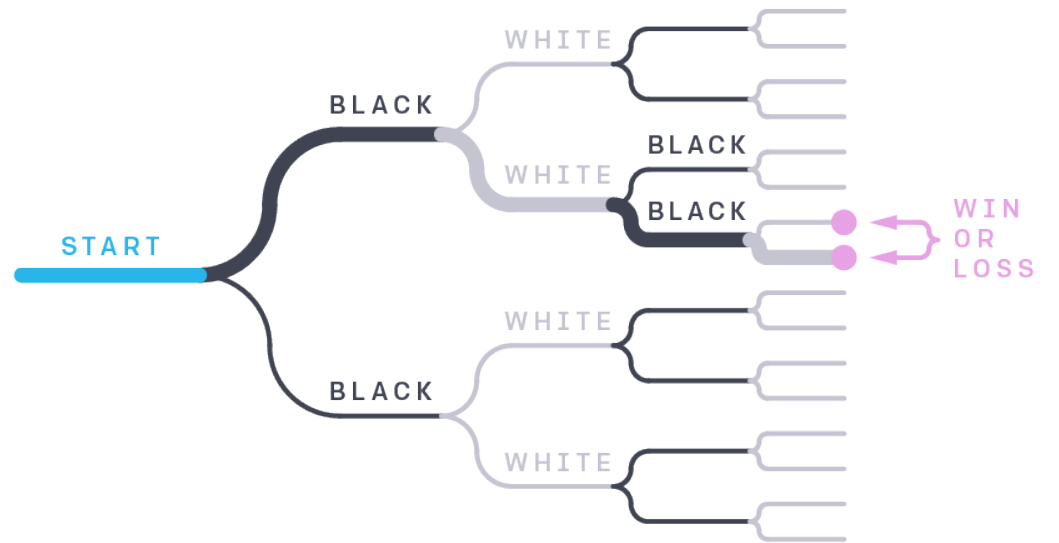


Debate

Tree of all possible debates



Tree of all possible Go moves



What are people doing about AGI safety?

- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI
 - Having the right goal (Ambitious value learning)
 - Human selection of behavior (IRL, Preference learning, Reward modeling)
 - Optimizing for *our* goals (Cooperative IRL)
 - Corrigibility (Iterated amplification, Debate, Factored cognition)

What are people doing about AGI safety?

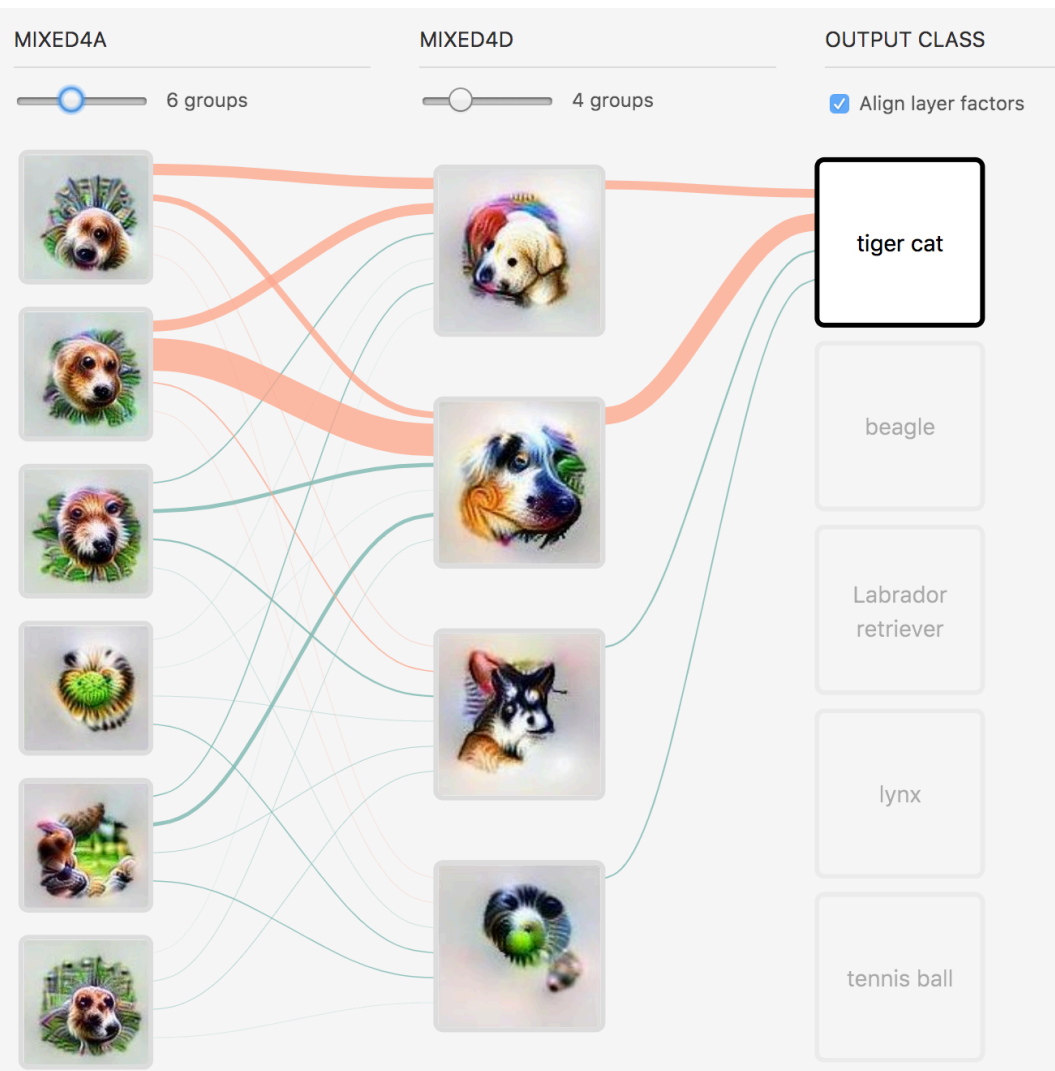
- What even is going on with AGI? (Embedded agency, CAIS)
- Limited AGI
 - Containment (AI Boxing)
 - Preventing bad behavior (Impact measures / Avoiding side effects)
- Robustness (Verification, Red teaming, Adversarial ML)
- Helpful AGI
 - Having the right goal (Ambitious value learning)
 - Human selection of behavior (IRL, Preference learning, Reward modeling)
 - Optimizing for *our* goals (Cooperative IRL)
 - Corrigibility (Iterated amplification, Debate, Factored cognition)
- Interpretability

Interpretability



To understand multiple layers together, we would like each layer's factorization to be "compatible"—to have the groups of earlier layers naturally compose into the groups of later layers. This is also something we can optimize the factorization for.

— positive influence
— negative influence



Takeaways

There are five main avenues of research: understanding AGI, limited AGI, robustness, helpful AGI, and interpretability.

We can try to build helpful AGI either by learning preferences and getting corrigibility as a result, or by learning corrigibility and getting preference learning as a result.

We can either try just to prevent catastrophic outcomes, or try to make the outcomes actively good.