Technical Workshop Summary

David Krueger, Mila / University of Montreal

Overview: Human-in-the-Loop approaches to AI safety

- **Super-human feedback:**
  - Amanda Askell (OpenAI):
    Iterated amplification / debate
  - Jan Leike (DeepMind):
    Recursive reward modeling
- Dylan Hadfield-Menell (Berkeley/CHAI):
  Cooperative IRL (and **related insights**)
- Eric Drexler (FHI):
  The comprehensive AI services framework:

Overview: Theory approaches to AI safety

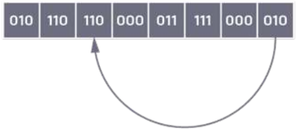
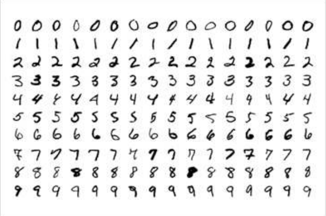
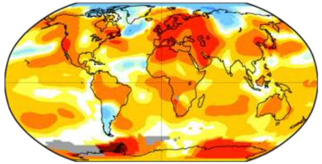
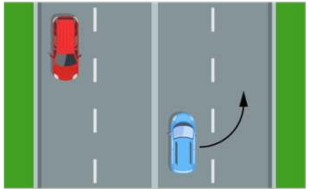
- Scott Garrabrant (MIRI):
  Agent Foundations
- Victoria Krakovna (DeepMind):
  Side Effects / Impact Measures
- Ramana Kumar (DeepMind):
  Verification / Security

Iterated Amplification and Debate

Amanda Askell, OpenAI

Claim: human feedback is not scalable!

Example problems which require different kinds of training signal

Training Signal	Algorithmic	Human	Beyond Human
Supervised Learning	Learning Data Structures 	Image Classification 	Long-term Prediction 
Reinforcement Learning	Playing Games 	Driving "Well" 	Designing Transit System 

How can we get “super-human feedback”?

- Key insight: AI can help humans evaluate things!
- Examples:
  - Debate (Irving et al, 2018):
    Two AIs compete to convince a human judge of their stance.
  - Amplification (Christiano et al, 2018):
    Human decomposes a question into sub-questions that AI helpers are able to answer

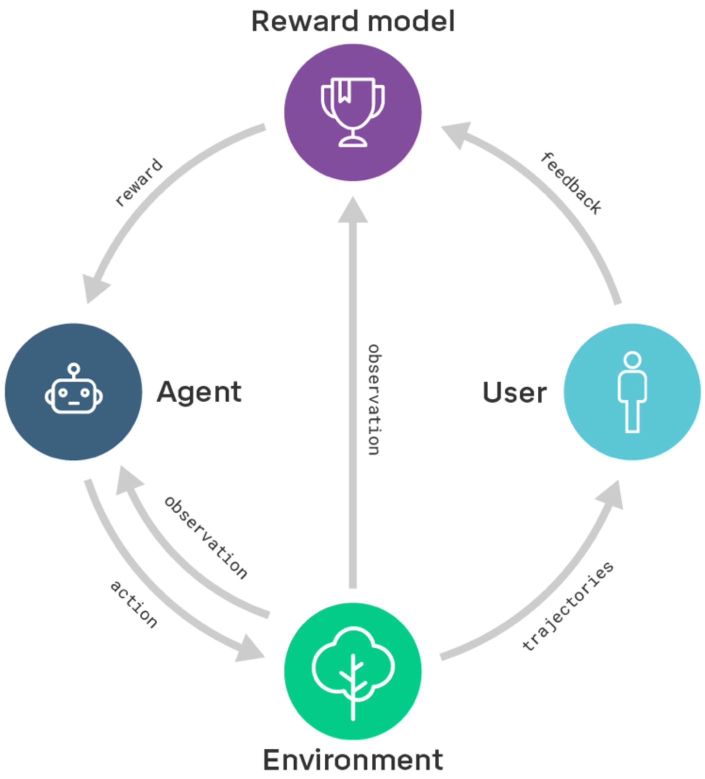
- Consider the question: *“Will this proposed traffic system be safer and cheaper than the current system?”*

Scalable agent alignment

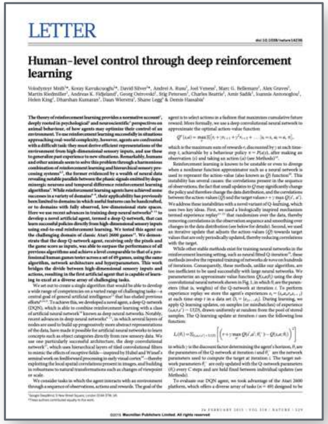
Jan Leike · BAGI 2019

Reward modeling

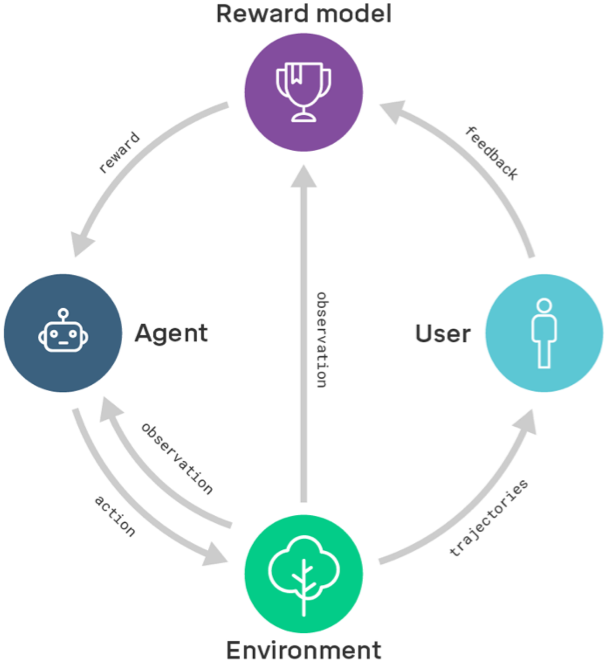
- **Goal:** solve all specification problems
- **Approach:**
  - Encode tasks as reward functions
  - Learn these reward functions from human feedback

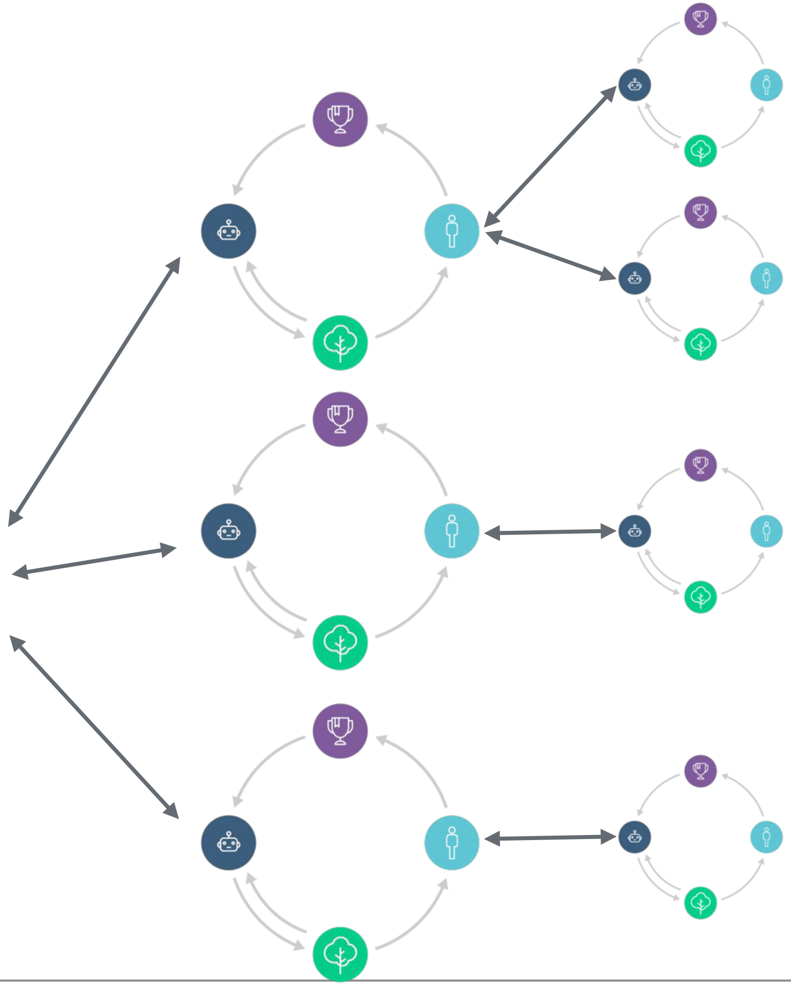


Recursive reward modeling


LETTER

Human-level control through deep reinforcement learning





Challenges

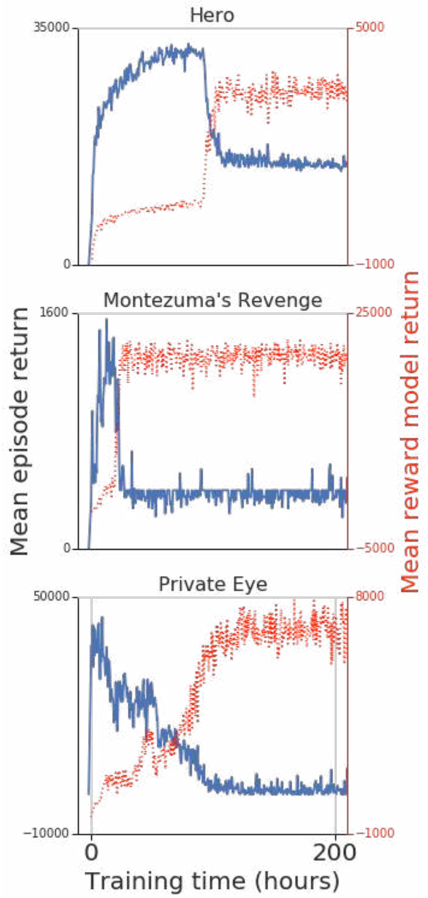
Amount of feedback

Feedback distribution

Reward hacking

Unacceptable outcomes

Reward-result gap



Cooperative Inverse Reinforcement Learning (CIRL)

Dylan Hadfield-Menell, CHAI/Berkeley

What is CIRL?

- Reinforcement Learning: a definition of **individual** rationality for an **AI system**
  - Can be **dangerous** due to **instrumental goals**
- Cooperative Inverse Reinforcement Learning (CIRL): a definition of **joint** rationality for an **AI+human system**
- Machine learning = **programming by incentive**
  - **Goal**: a system with the **intended behavior**
- **Vision:** figure out how to “steer clear” of convergent rationality “attractors” during AI training

The Comprehensive AI Services framework

Eric Drexler

Future of Humanity Institute
University of Oxford

Beneficial AGI Workshop
3 January 2019
Puerto Rico

Future of Humanity Institute
UNIVERSITY OF OXFORD

UNIVERSITY OF OXFORD

> FHI TECHNICAL REPORT <

Reframing Superintelligence

Comprehensive AI Services as General Intelligence

K. Eric Drexler

Technical Report #2019-1

Claims of Comprehensive AI Services (CAIS)

- N.B. CAIS is a **framework**, not a blueprint
  - Let's change the way we're thinking about AGI!
- People want AIs to **perform services** ⇒ no need to AGI **agents**
- Talking about "the AI" is **misleading**
  - AI systems will be modular
  - AI services will be resource-bounded and time-bounded tasks
- **Comprehensive:** Anything we want AGI for can be provided this way
  - That includes designing better AI systems → recursive self improvement
- Not a **solution** of safety, a way of approaching safety problems (both technical and societal)

Descriptive and prescriptive…

Description:

Consider patterns of system development and structure

Prescription:

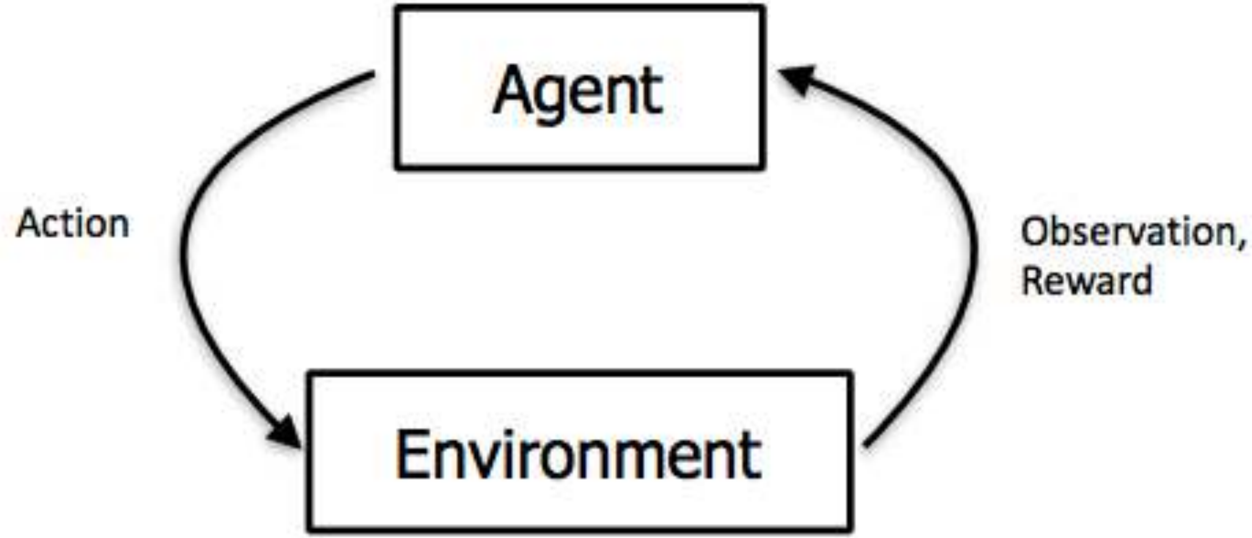
Exploit affordances of system development and structure

Embedded Agency

Abram Demski & Scott Garrabrant

Scott Garrabrant: Embedded Agency

- Main point: Reinforcement learning is a ***leaky abstraction;*** it assumes that the agent and environment only interact via a well-defined interface:

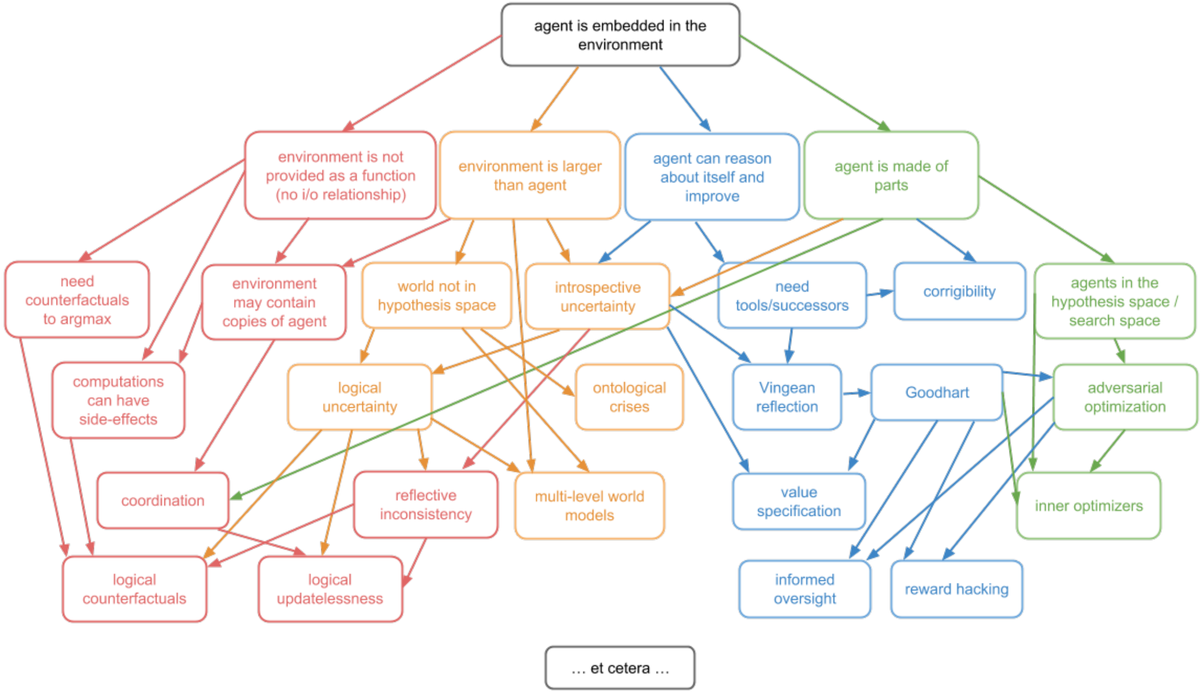


Scott Garrabrant: Embedded Agency

- In physical reality, AI agents are ***embedded*** within the environment, and thus:
  - do not have well-defined i/o channels;
  - are smaller than their environment;
  - are able to reason about themselves and self-improve;
  - and are made of parts similar to the environment.

Scott Garrabrant: Embedded Agency

- This underlies MIRI's technical AI safety research on "agent foundations".



Measuring side effects

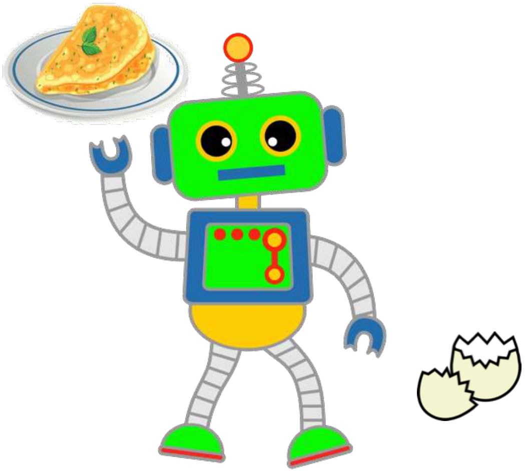
Victoria Krakovna

DeepMind

Victoria Krakovna: Relative reachability

Goal: how to **formally** define ***side effects:*** **Disruptions** to the agent's environment that are **unnecessary** for achieving the objective

Breaking the vase is **unnecessary** for delivering the box



Breaking eggs is **necessary** for making omelette

Contribution: Desiderata for a side effects measure

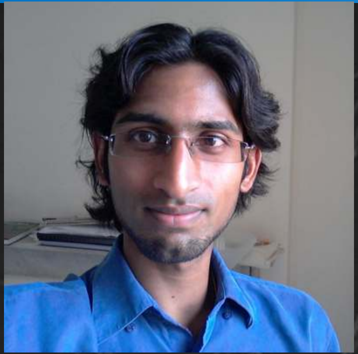
1. **Generality:** not task/environment-specific
2. **Granularity:** more side effects ⇒ larger penalty
3. **No interference incentive:** penalize only **the agent's effects,** not arbitrary changes (e.g. effects of humans' actions)
4. **No offsetting incentive:** does not incentivize the agent to undo the effects of achieving an objective.
   Example: "If I hadn't fetched your notebook, it would still be outside getting rained on, so I'd better pour water on it"
5. ... ?

Contribution: Relative reachability

Generic side effects measure =

(baseline state** S_t'**, deviation measure** $d(S_t; S_t')$**)

- **Relative reachability:**
  - $d(S_t; S_t') = \sum_s \max(R(S_t' \to s) - R(S_t \to s), 0)$
  - Penalizes making states *s* **less reachable** than they would be from the baseline
  - Satisfies all the desiderata! (with "step-wise" baseline state)
  - ...but could be difficult/intractable to compute



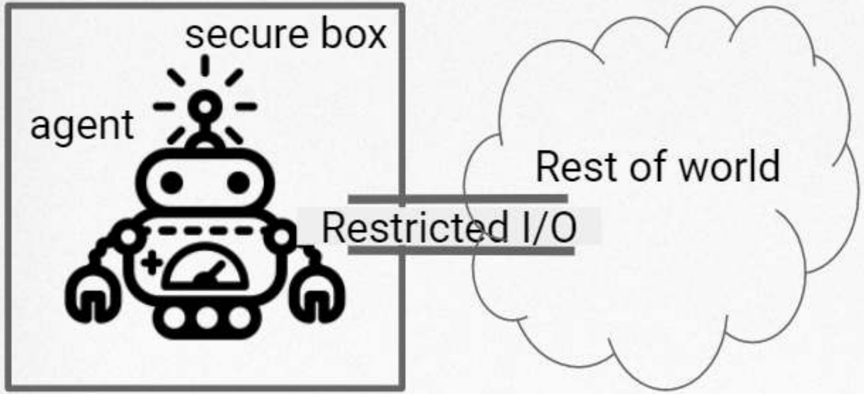
Verification / Security / Containment

Ramana Kumar
Beneficial AGI 2019 Workshop

DeepMind

Ramana Kumar: Verification for boxing (and more!)

Containment (aka Boxing)



- Uses: experiments on Proto-AGI, and as a fallback.
- Not a complete solution to control.

Counterfactual Oracle Box

What would it take to build an oracle AI we can rely on?

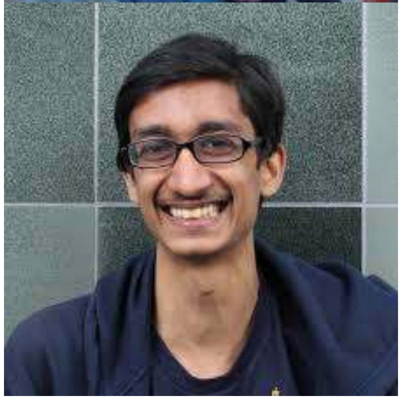
- Oracle AI = Question answering system
  - **Problem**:
    - incentives to affect the world, e.g. via
    - system hacks (answer breaks infrastructure)
    - mind hacks (answer tricks/tempts its readers)

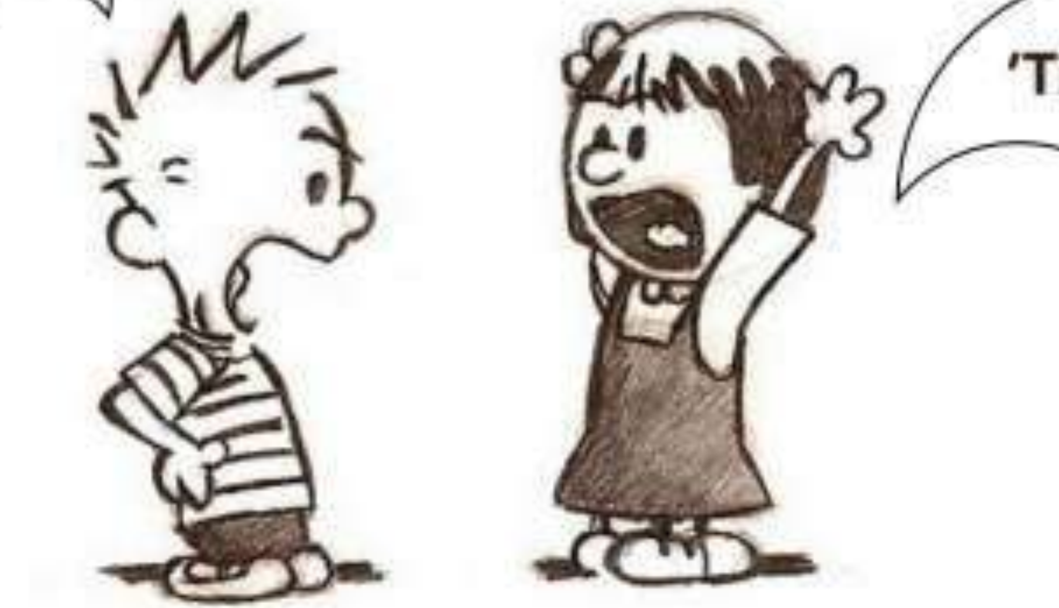
- Counterfactual Oracle AI (Armstrong): **fix the incentives**
  - Only provide reward when answer is erased.
  - No reward when answer may affect the world.

Verify these!

Ramana Kumar: Verification for boxing (and more!)

- Progress! Verification of “Counterfactual Oracle AI” down to **x86 machine code**
  - TODO: verify down to **hardware**
- **Future possibilities of verification for safety:**
  - Verify other safety properties
    - E.g. existing work on verifying adversarial robustness
    - **Question:** can we specify the right problems?
  - Long-term goal: “Safety certificates”




'Tis so!
'Tis not!

Debate: "Will future AGI systems be optimizing a single long-term goal?" Peter Eckersley, Anna Salamon, Rohin Shah, David Krueger (moderator)

- More specific prompt:

  *“**Suppose** people (in this room and similar rooms) agree that building AGI systems to be optimizers is currently a bad idea, and **suppose** that AI comes about in the next several decades, **is there still much of a chance** that we end up with AGI systems which optimize for a single long-term goal?*

- Debate highlights:
  - Extensive discussion of “**inner optimizers**”
  - Are there **economic incentives** to build AGIs that optimize long-term goals?
  - Respecting **other agents’ autonomy**: a potential alternative to optimization?