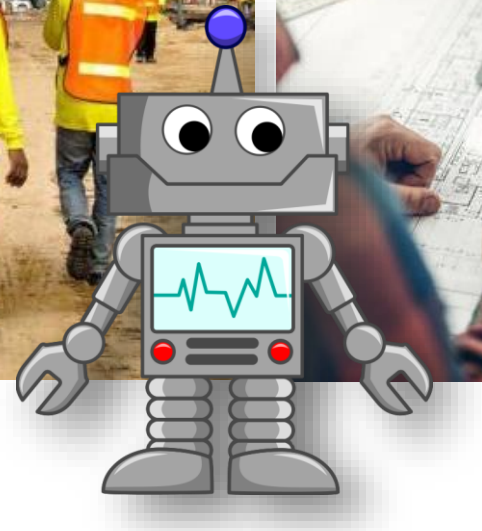
High-Confidence Guarantees for Safe Reward and Policy Learning Under Uncertainty

Daniel Brown

KAHLERT SCHOOL OF COMPUTING
THE UNIVERSITY OF UTAH

ROBOTICS CENTER
THE UNIVERSITY OF UTAH



Objectives are often very hard to specify!

The Alignment Problem

How do we get AI systems to do what we, as humans, actually want them to do?

Robust behavior is often very hard to specify!

How do we know if an AI system is robust?

Robustness: Acceptable behavior in the presence of uncertainty and unusual circumstances.

How do we know if an AI system is robust?

Robustness: **Acceptable behavior** in the presence of uncertainty and unusual circumstances.

The Alignment Problem

Human input is messy!

Aligned, Robust, and Interactive Autonomy (ARIA) Lab

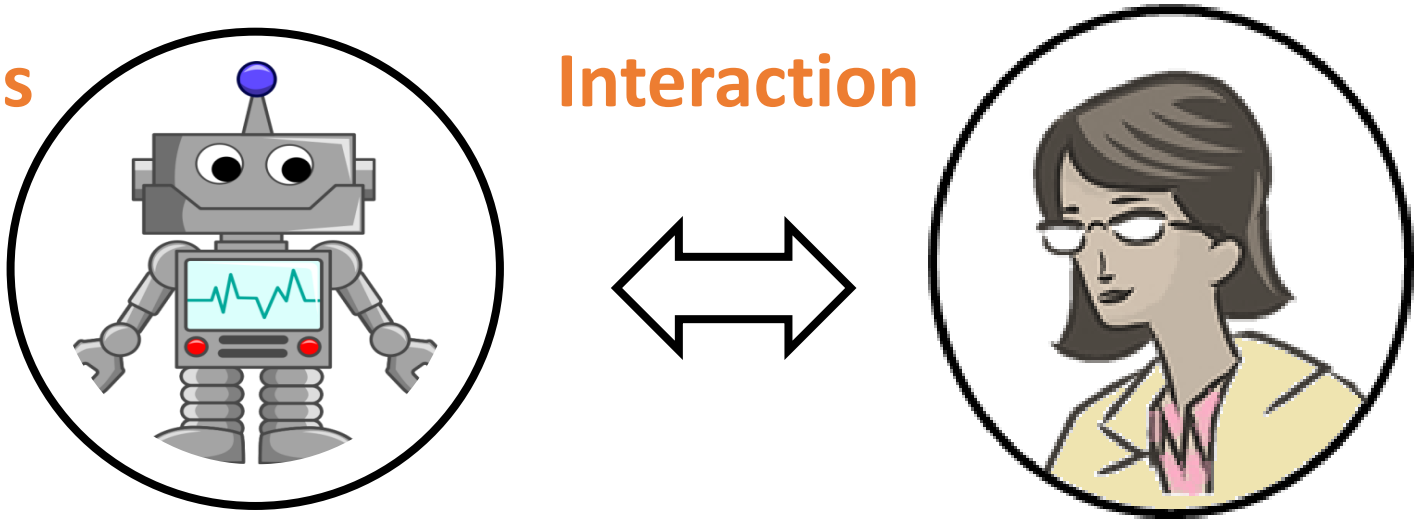
Our research seeks to efficiently incorporate human input into both the theory and practice of robust and aligned AI systems.

The Alignment Problem

Robustness

Interaction

Irrationality

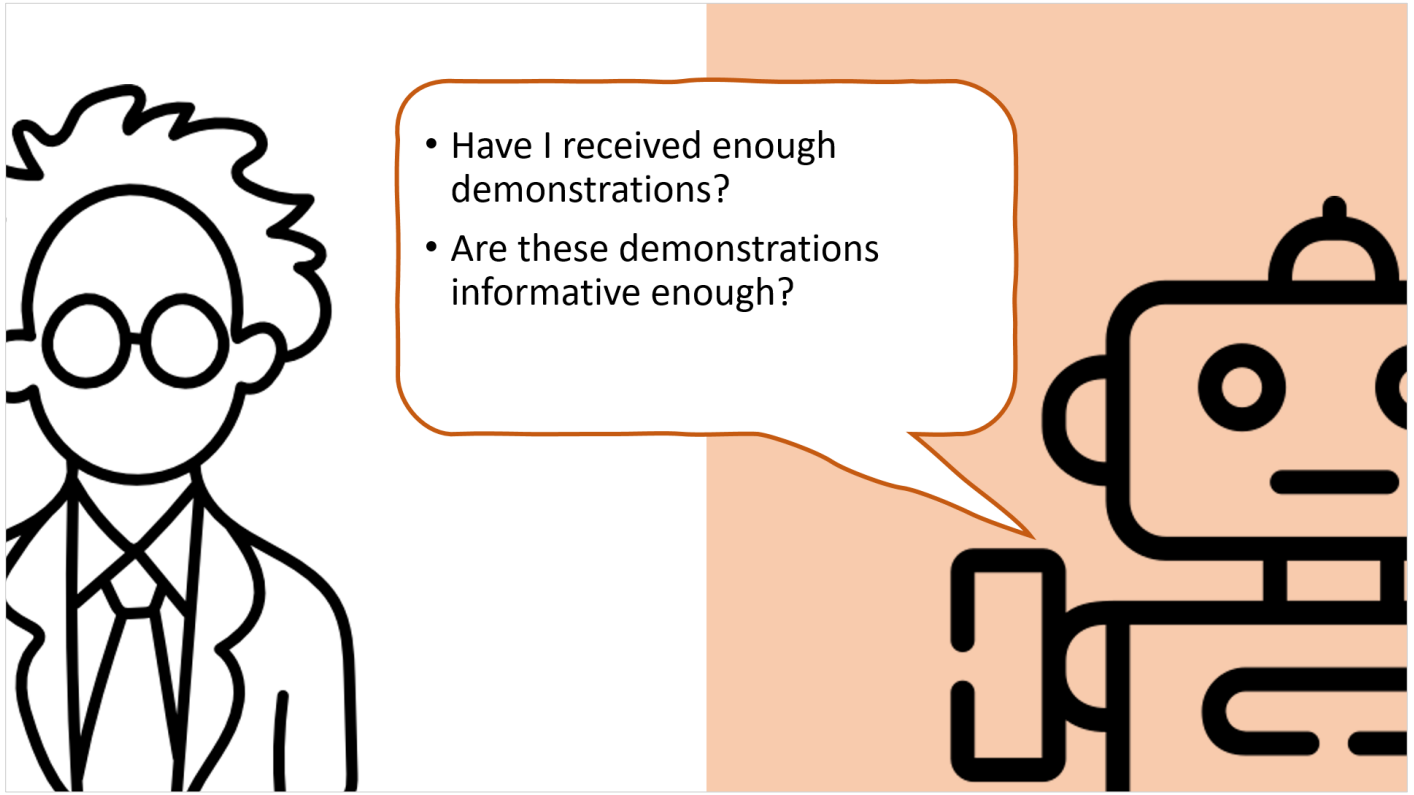


Uncertainty

Feedback

Autonomous Assessment of Demonstration Sufficiency via Bayesian Inverse Reinforcement Learning

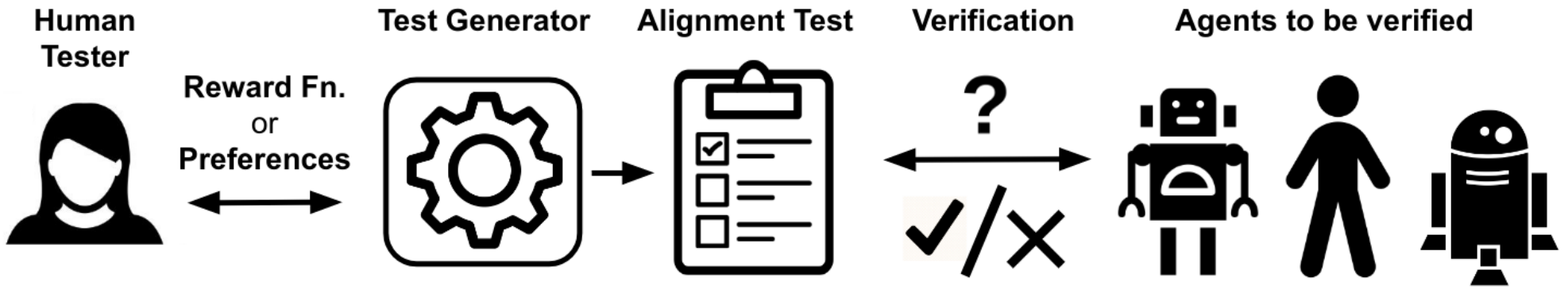
Tu (Alina) Trinh Haoyu Chen Daniel S. Brown



Best paper finalist award at International Conference on Human-Robot Interaction (HRI'24)!

Research Goal 1:
Probabilistic performance bounds when learning rewards from any type of human feedback.

Value Alignment Verification for AI systems



Brown et al. "Value Alignment Verification." *ICML*, 2021.

Research Goal 2:
Unit Tests for Reward and Policy Alignment.

Goal Misgeneralization in Deep Reinforcement Learning

Lauro Langosco [*1] **Jack Koch** [*] **Lee Sharkey** [*2] **Jacob Pfau** [3] **Laurent Orseau** [4] **David Krueger** [1]

Abstract

We study *goal misgeneralization*, a type of o
of-distribution generalization failure in reinfor
ment learning (RL). Goal misgeneralization
curs when an RL agent retains its capabilities o
of-distribution yet pursues the wrong goal. F
instance, an agent might continue to competen
avoid obstacles, but navigate to the wrong pla

Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals

Rohin Shah [* †] **V**
rohinmshah@deepmind.com vikrant

Mary Phuong [†] **Victoria Krakovna**

CAUSAL CONFUSION AND REWARD MISIDENTIFICATION IN PREFERENCE-BASED REWARD LEARNING

Jeremy Tien
University of California, Berkeley
jtien@berkeley.edu

Jerry Zhi-Yang He
University of California, Berkeley

Zackory Erickson
Carnegie Mellon University

Anca D. Dragan
University of California, Berkeley

Daniel S. Brown
University of Utah

Research Goal 3:
Robust Policy Optimization Under Reward Uncertainty

High-Confidence Guarantees for Safe Reward and Policy Learning Under Uncertainty

Daniel Brown

KAHLERT SCHOOL OF COMPUTING
THE UNIVERSITY OF UTAH

ROBOTICS CENTER
THE UNIVERSITY OF UTAH