Tool-Use Robot Manipulation Tasks for Cooperative and Explainable Operations in Safety-Critical Domains

Dissertation Defense, March 2025

Emily Sheetz

Advised by Benjamin Kuipers

Committee: Joyce Chai, Kimberly Hambuchen, and Chad Jenkins

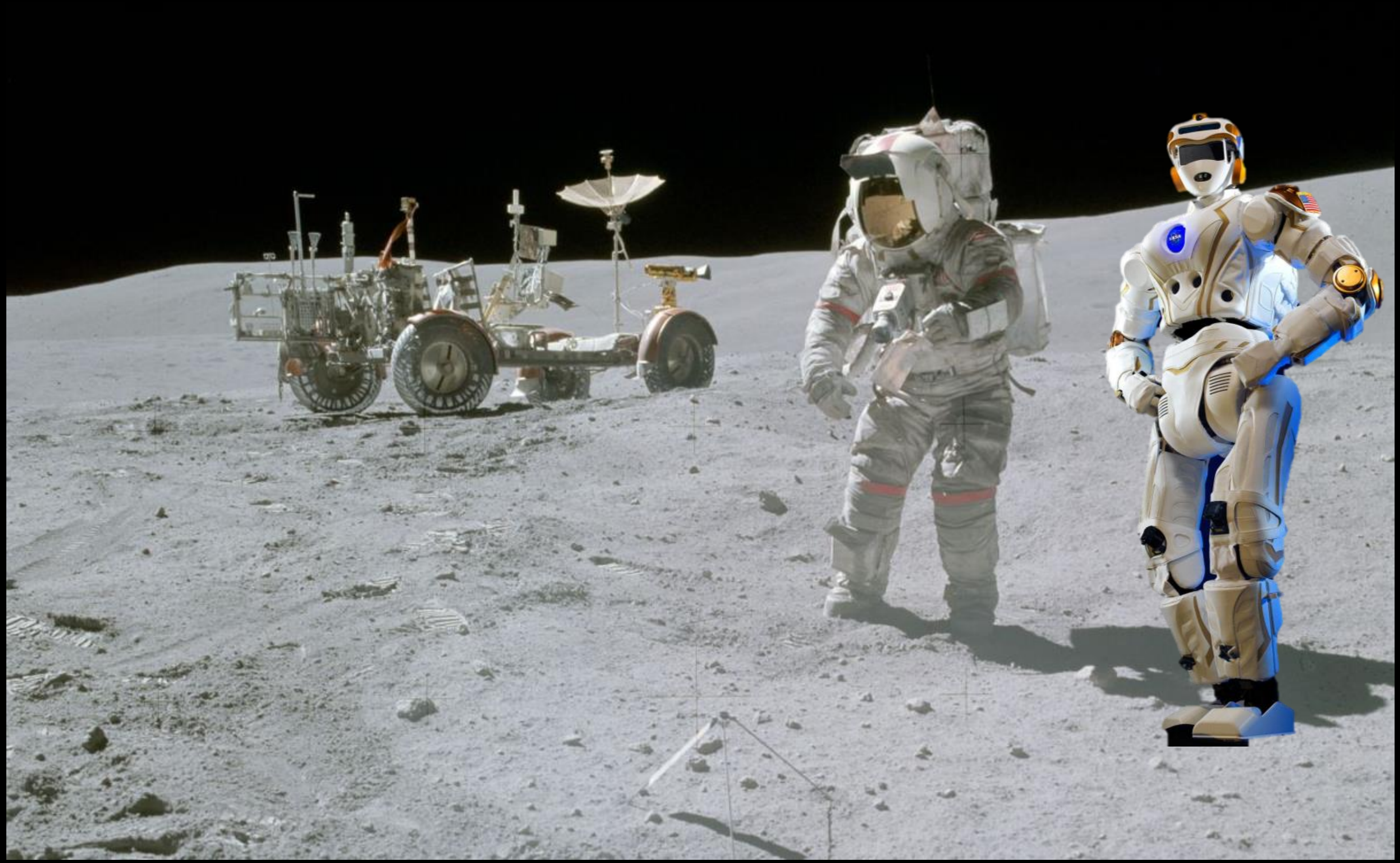
This work was supported by a NASA Space Technology Graduate Research Opportunity (NSTGRO).

Problem and Motivation

Robotics for Space Exploration

The Artemis missions will return humanity to the Moon to learn about establishing continuous presence in space.

Crew time could be used more effectively for science mission goals if robots help with the many tasks of space exploration.



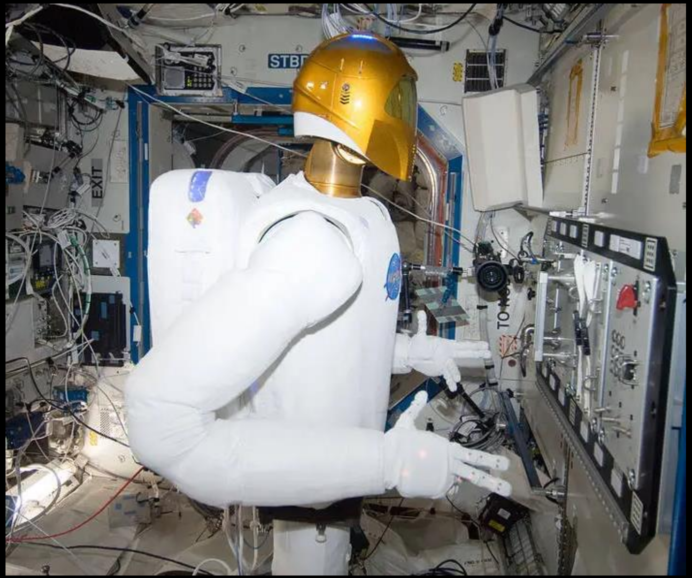
Inspired by slides from the Johnson Space Center Dexterous Robotics Team.

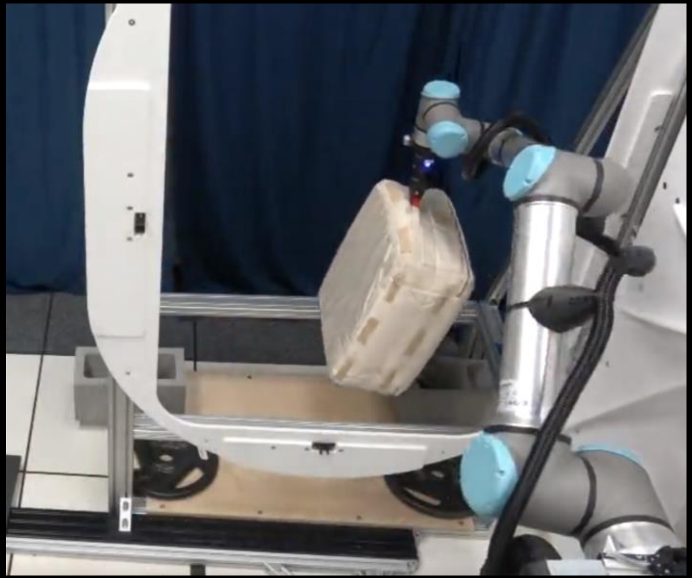
[1] NASA, "Artemis," *NASA*, [Online], 2024.
[2] NASA, "Moon to Mars Architecture," *NASA*, [Online], 2024.

Human-Robot Teams in Extreme Domains

We need robots to operate as capable, trusted agents on human-robot teams in various safety-critical problem domains.

To achieve this, we consider two key challenges: (1) robot manipulation capabilities and (2) robot safety reasoning.



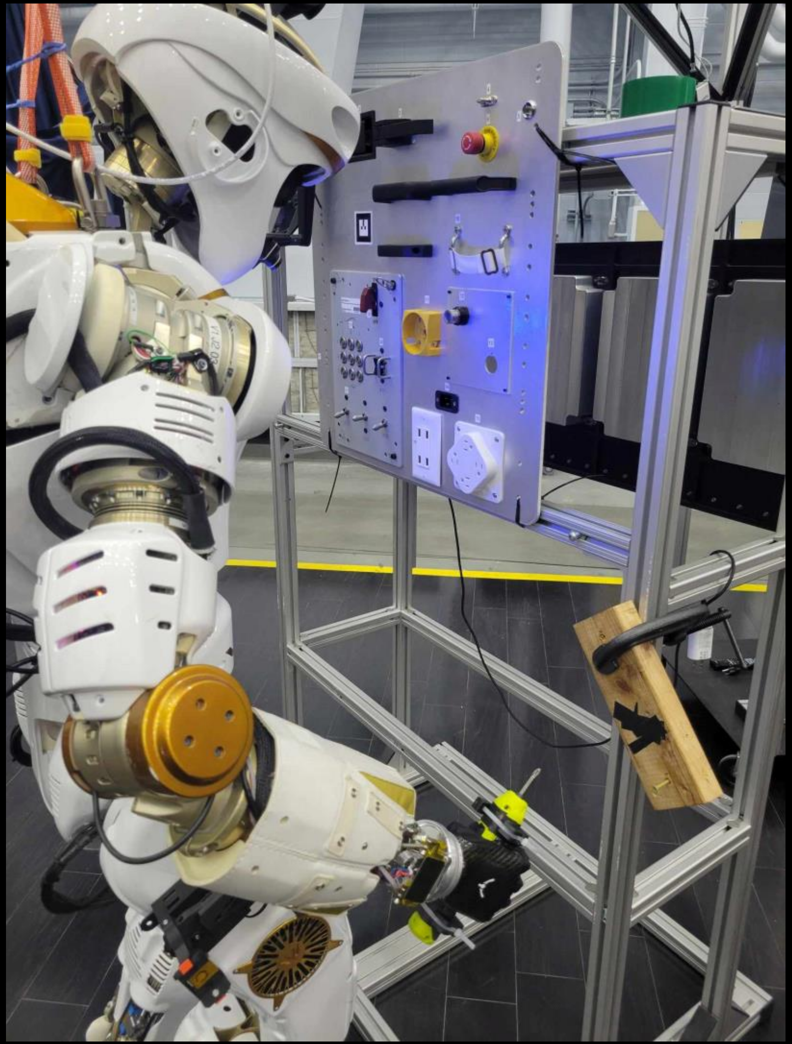


Tool-Use Tasks on Human-Robot Teams

Reasoning over *object affordances* ("action possibilities" or "opportunities for action") and executing afforded actions in tool-use tasks are challenging.

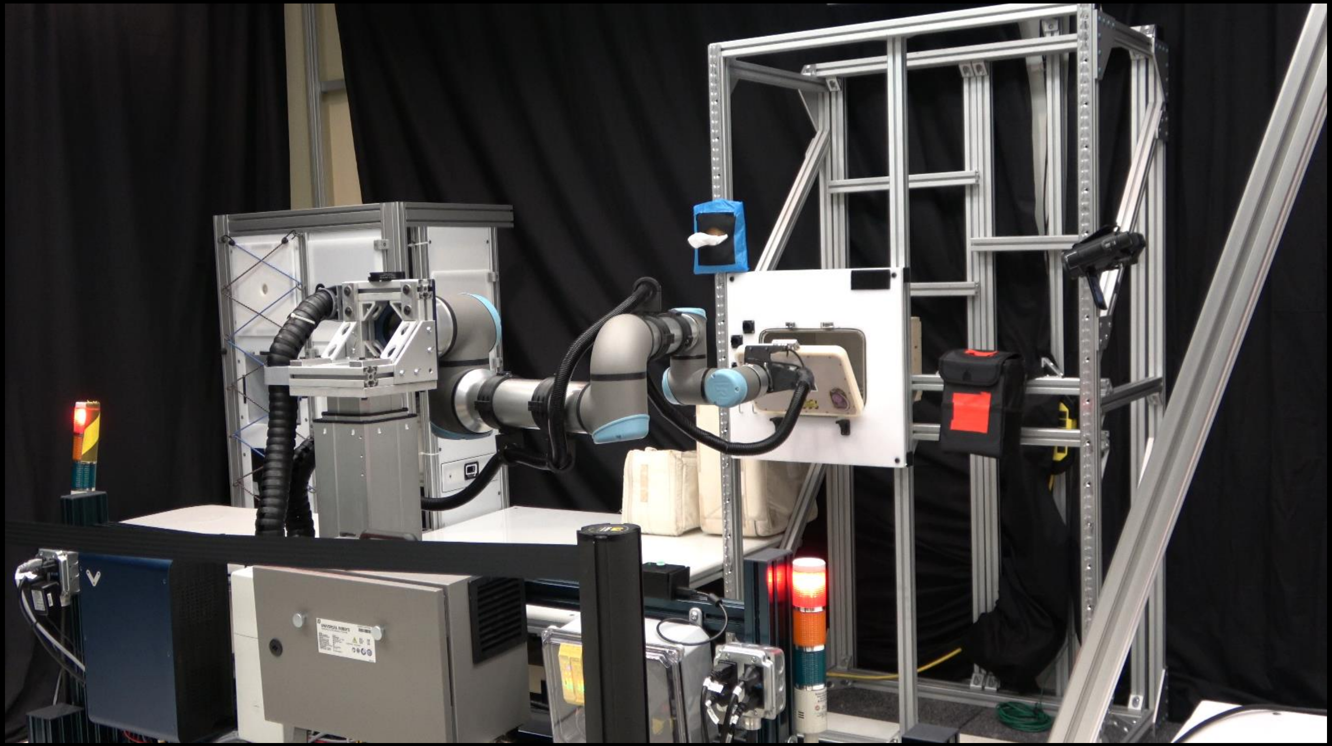
We want robots to function on human-robot teams alongside humans without robotics experience.

Therefore, we want our methods and results to be explainable and minimize expert knowledge engineering.



[3] J. Gibson, "The Theory of Affordances," *Perceiving, Acting, and Knowing: Towards an Ecological Psychology*, 1977.
[4] AMP von Bayern *et al.*, "Compound Tool Construction by New Caledonian Crows," *Scientific Reports*, 2018.

Trust and Safety on Human-Robot Teams



Safety and trust are important when robots operate alongside humans.

Human operators may mistrust or overtrust robots when expectations do not align with the robot's capabilities.

We explore explainable methods to promote understanding and trust for safe operations on human-robot teams.

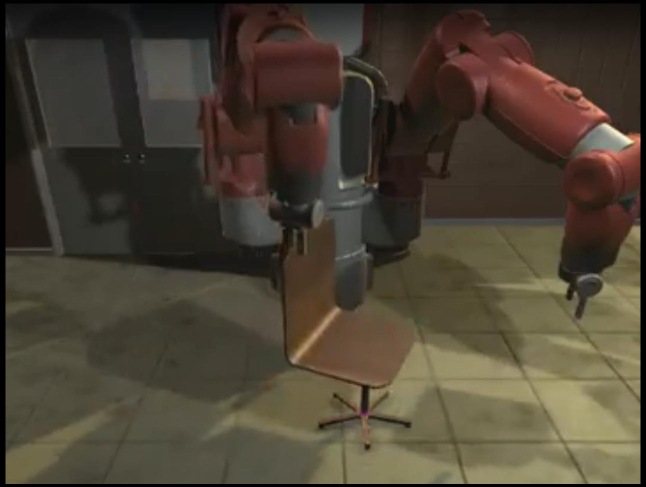
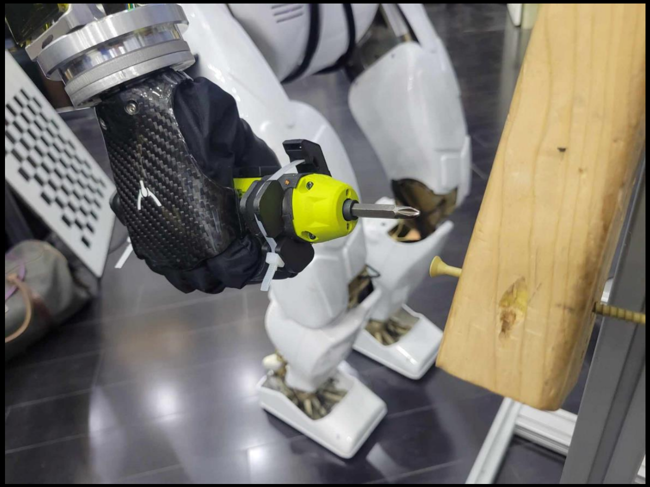
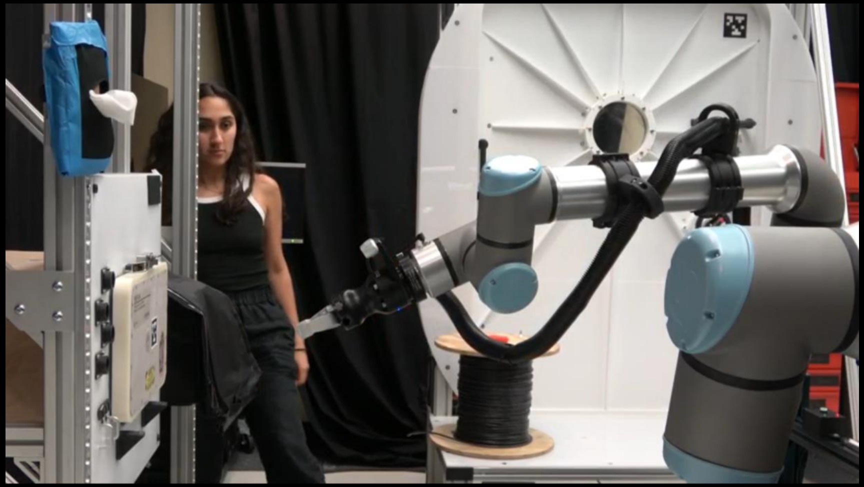
[5] M. Vasic and A. Billard, "Safety Issues in Human-Robot Interactions," *IEEE ICRA*, 2013.
[6] Y. Zhang *et al.*, "DANLI: Deliberative Agent for Following Natural Language Instructions," *arXiv preprint arXiv:2210.12485*, 2022.
[7] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, 2004.
[8] P. Robinette *et al.*, "Overtrust of Robots in Emergency Evacuation Scenarios," *IEEE Conference on HRI*, 2016.
[9] B. Kuipers, "Trust and Cooperation," *Frontiers in Robotics and AI*, 2022.

Dissertation Contributions

Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

Safety reasoning on human-robot teams
(UR 2025, Under Review)

Dissertation Contributions

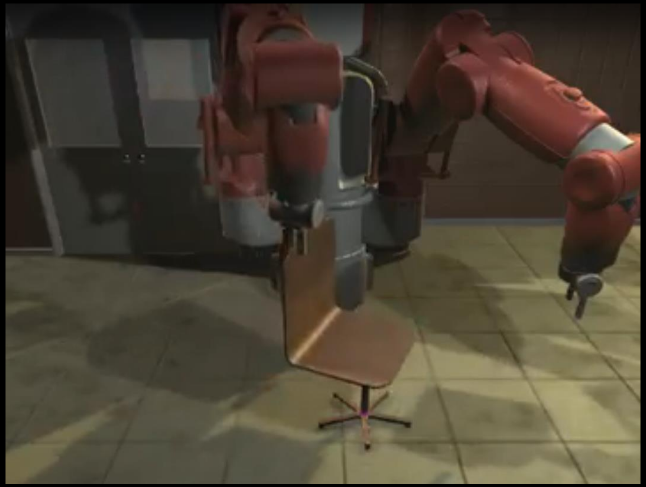
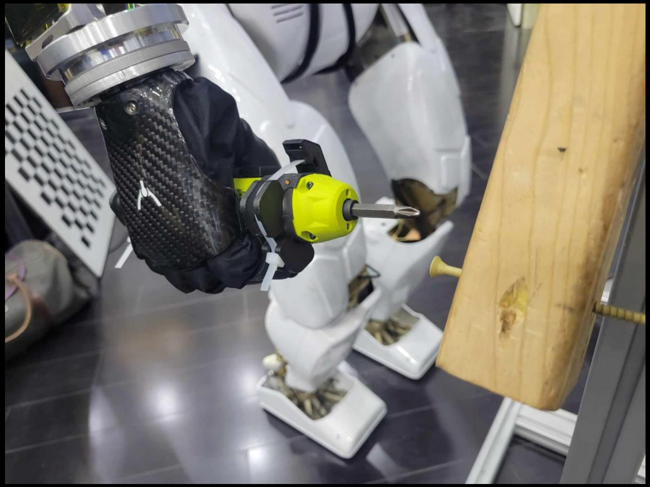
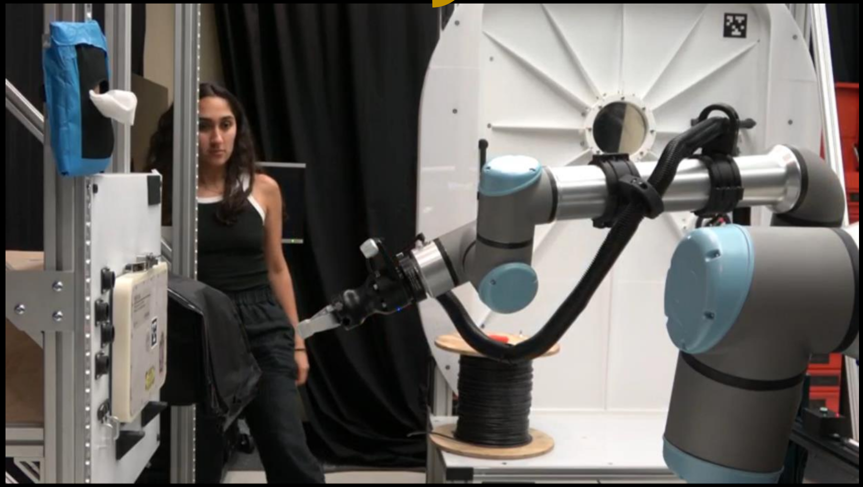
(higher-level discussion)

Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

Safety reasoning on human-robot teams
(UR 2025, Under Review)

(most technical detail)

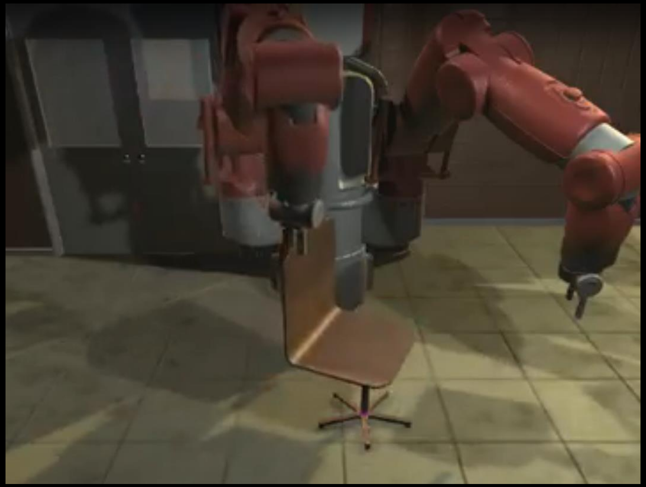




Dissertation Contributions

Autonomous planning of complex assembly actions
(ICRA 2022)

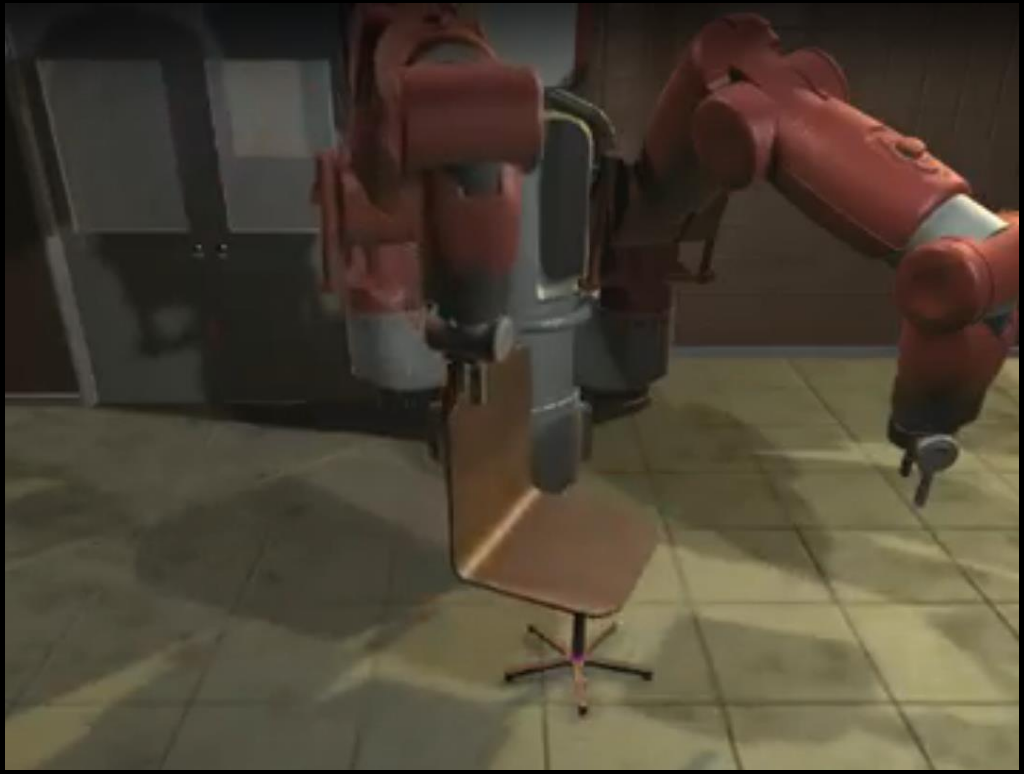
Reliable and explainable execution of tool-use tasks
(IROS 2024)

Safety reasoning on human-robot teams
(UR 2025, Under Review)



Planning Complex Actions: Causal Control Basis

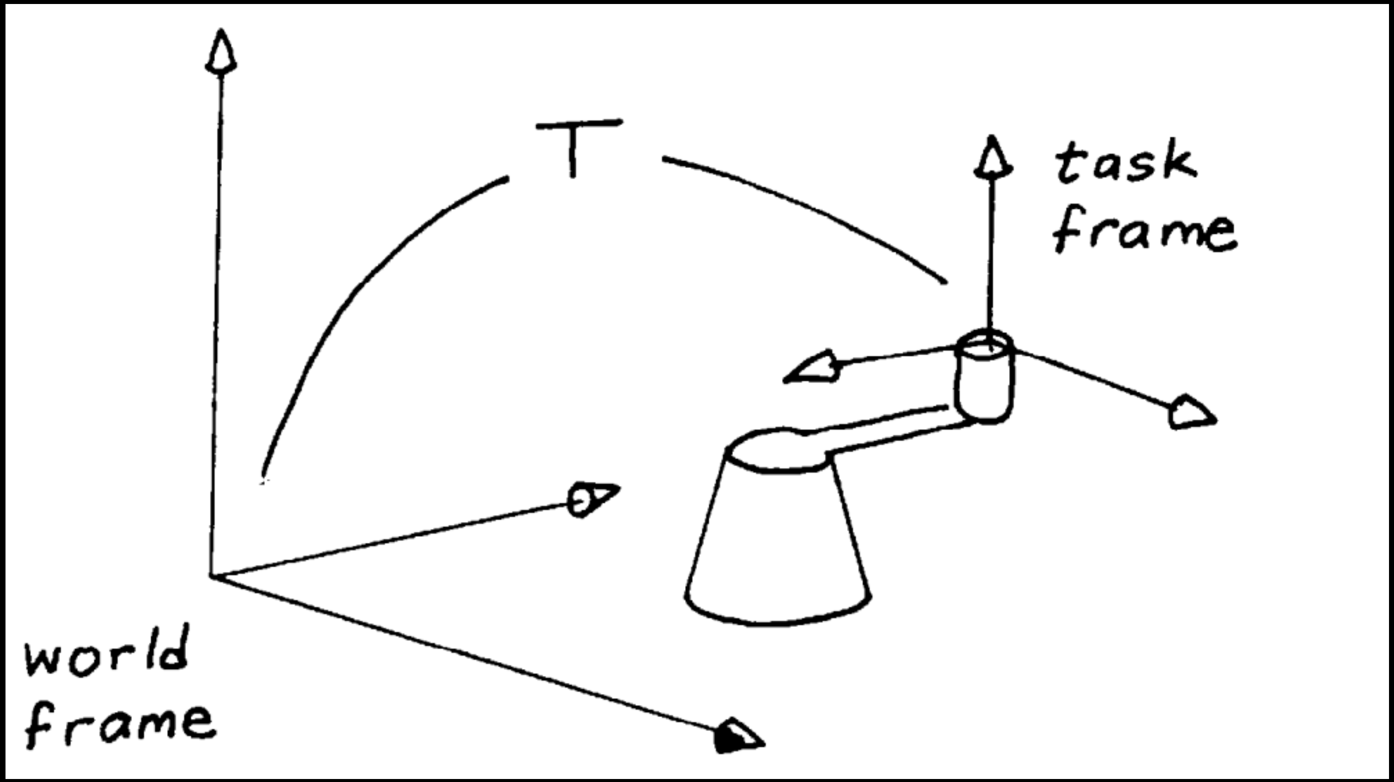
Complex Actions in Assembly Tasks



Tool-use and assembly tasks require advanced planning over objects, their separate parts, and their affordances.

[10] J. Pavlasek, S. Lewis, K. Desingh, and O. C. Jenkins, “Parts-Based Articulated Object Localization in Clutter Using Belief Propagation,” *IEEE IROS*, 2020.

Object-Centric Controllers



Reasoning over affordances and executing actions can be simplified with object-centric controllers, which formulate objectives with respect to object or task frames instead of the world frame.

[11] D. H. Ballard, "Task Frames in Robot Manipulation," *AAAI*, 1984.
[12] S. Hart, P. Dinh, and K. Hambuchen, "The Affordance Template ROS Package for Robot Task Programming," *IEEE ICRA*, 2015.

Controller Compositions

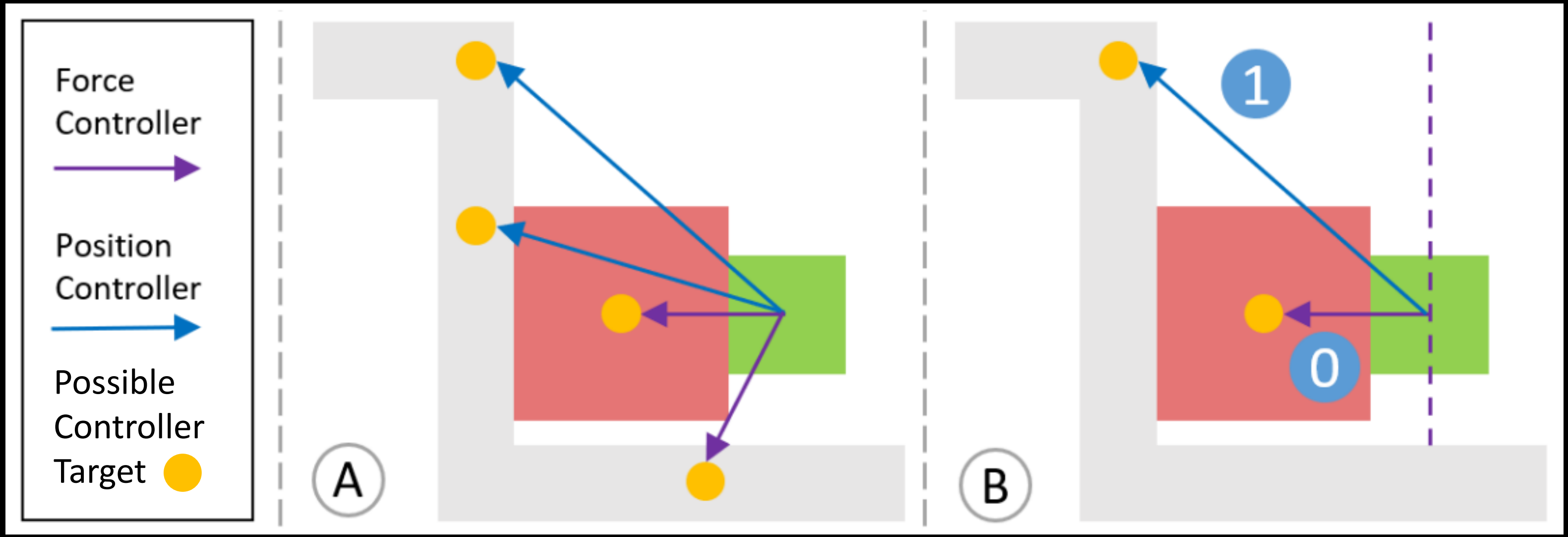
Object-centric controllers can be composed to create a behavior with a multi-objective role in a plan, meaning it achieves multiple task goals.

Composed behaviors prioritize one behavior over another so they can be executed concurrently.

[13] R. Platt, A. H. Fagg, R. A. Grupen, "Whole Body Grasping," [Online], 2004.
[14] R. Platt, A. H. Fagg, R. A. Grupen, "Nullspace Composition of Control Laws for Grasping," *IEEE IROS*, 2002.
[15] R. Platt, A. H. Fagg, R. A. Grupen, "Manipulation Gaits: Sequences of Grasp Control Tasks," *IEEE ICRA*, 2004.
[16] R. Platt, A. H. Fagg, R. A. Grupen, "Null-Space Grasp Control: Theory and Experiments," *IEEE Transactions on Robotics*, 2010.

Challenge: Autonomous Composition

Many works compose controllers using predefined priorities based on expert experience. We want the robot to autonomously compose controllers to minimize expert knowledge engineering.



For example, the robot needs to use its gripper (green block) to push the red block up the grey wall. We expect the robot to autonomously prioritize the given controllers. In this case, it prioritizes force (0) over positioning (1).

[14] R. Platt, A. H. Fagg, R. A. Grupen, "Nullspace Composition of Control Laws for Grasping," *IEEE IROS*, 2002.
[15] R. Platt, A. H. Fagg, R. A. Grupen, "Manipulation Gaits: Sequences of Grasp Control Tasks," *IEEE ICRA*, 2004.
[17] M. Sharma *et al.*, "Hierarchical Object-Centric Controllers for Robotics Manipulation," *arXiv preprint arXiv:2011.04627*, 2020.
[18] S. Hart and R. Grupen, "Natural Task Decomposition with Intrinsic Potential Fields," *IEEE IROS*, 2007.

Insight: Causality

We take inspiration from causal reasoning, or cause-effect relationships in long-horizon tasks.

We expect the robot to autonomously compose controllers by quantitatively predicting which composed behavior will achieve the intended composed effects within a task plan.

[19] C. Xiong *et al.*, "Robot Learning with a Spatial, Temporal, and Causal And-Or Graph," *IEEE ICRA*, 2016.
[20] J. Pearl, *Causality*, Cambridge University Press, 2009.
[21] I. Dasgupta *et al.*, "Causal Reasoning from Meta-Reinforcement Learning," *arXiv preprint arXiv:1901.08162*, 2019.

Causal Control Basis

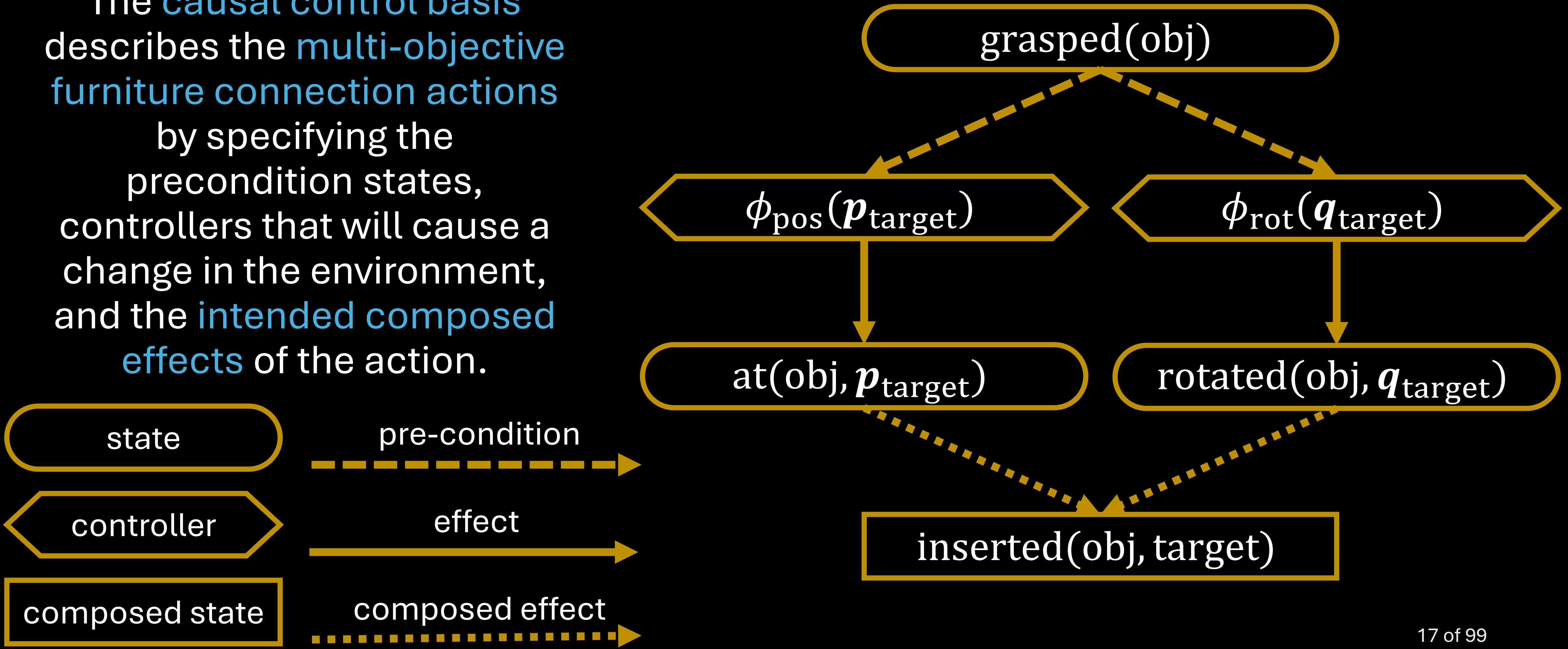
We propose a causal control basis, which annotates a control basis (set of object-centric controllers that form the building blocks of action execution) with causal graphs to enable autonomous controller compositions based on the intended composed effects.

We test our approach in furniture assembly tasks.

[19] C. Xiong *et al.*, "Robot Learning with a Spatial, Temporal, and Causal And-Or Graph," *IEEE ICRA*, 2016.

Composed Causal Graphs

The causal control basis describes the multi-objective furniture connection actions by specifying the precondition states, controllers that will cause a change in the environment, and the intended composed effects of the action.

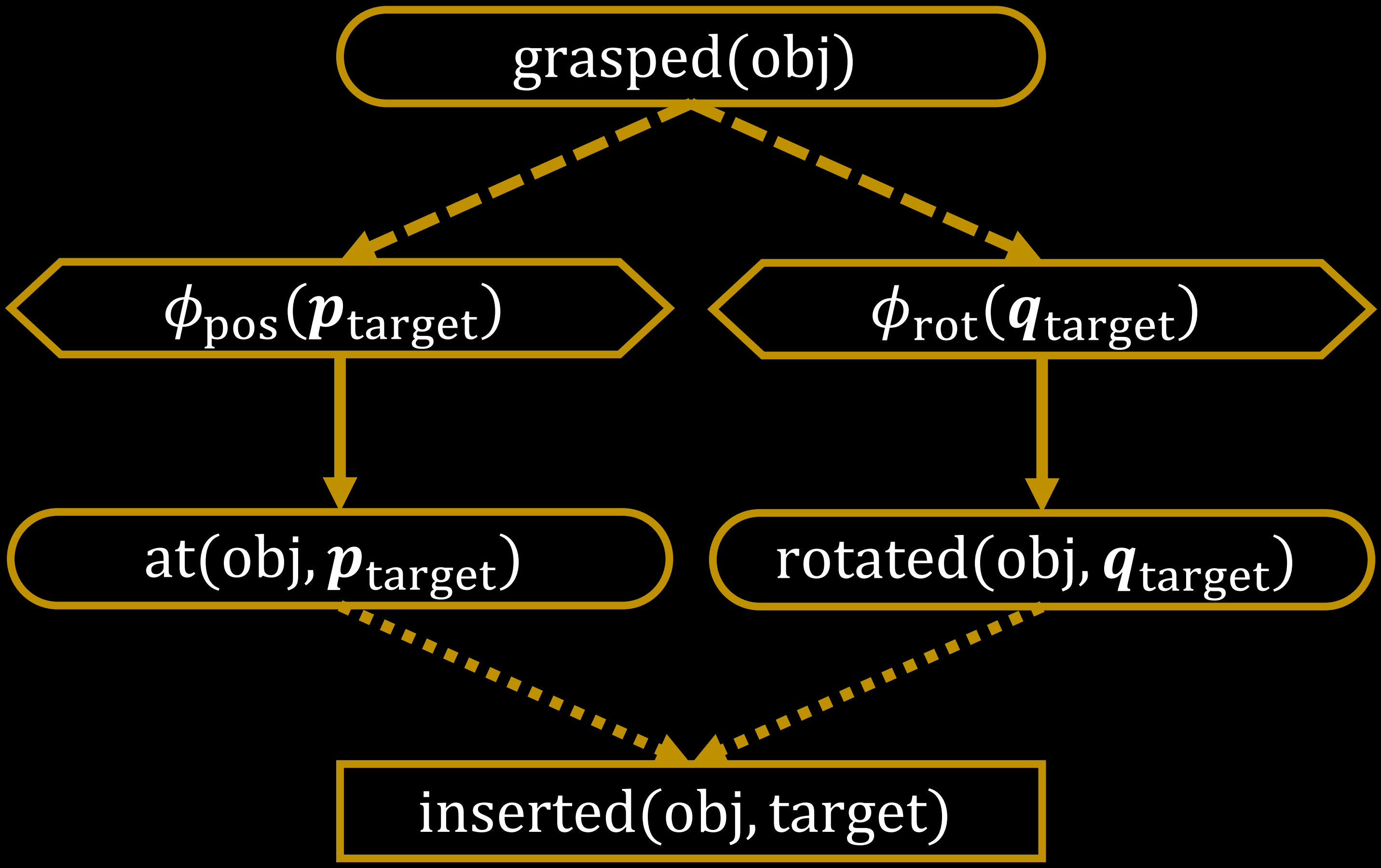


Composed Causal Graphs

The composed causal graph indicates what behaviors may achieve the composed effects, but not how to compose the controllers to realize these effects.

$$\phi_{\text{pos}} \lhd \phi_{\text{rot}}$$

$$\underbrace{\phi_{\text{rot}}}_{\substack{\text{left-hand side:}\\ \text{lower-priority}}} \lhd \underbrace{\phi_{\text{pos}}}_{\substack{\text{right-hand side:}\\ \text{higher-priority}}}$$

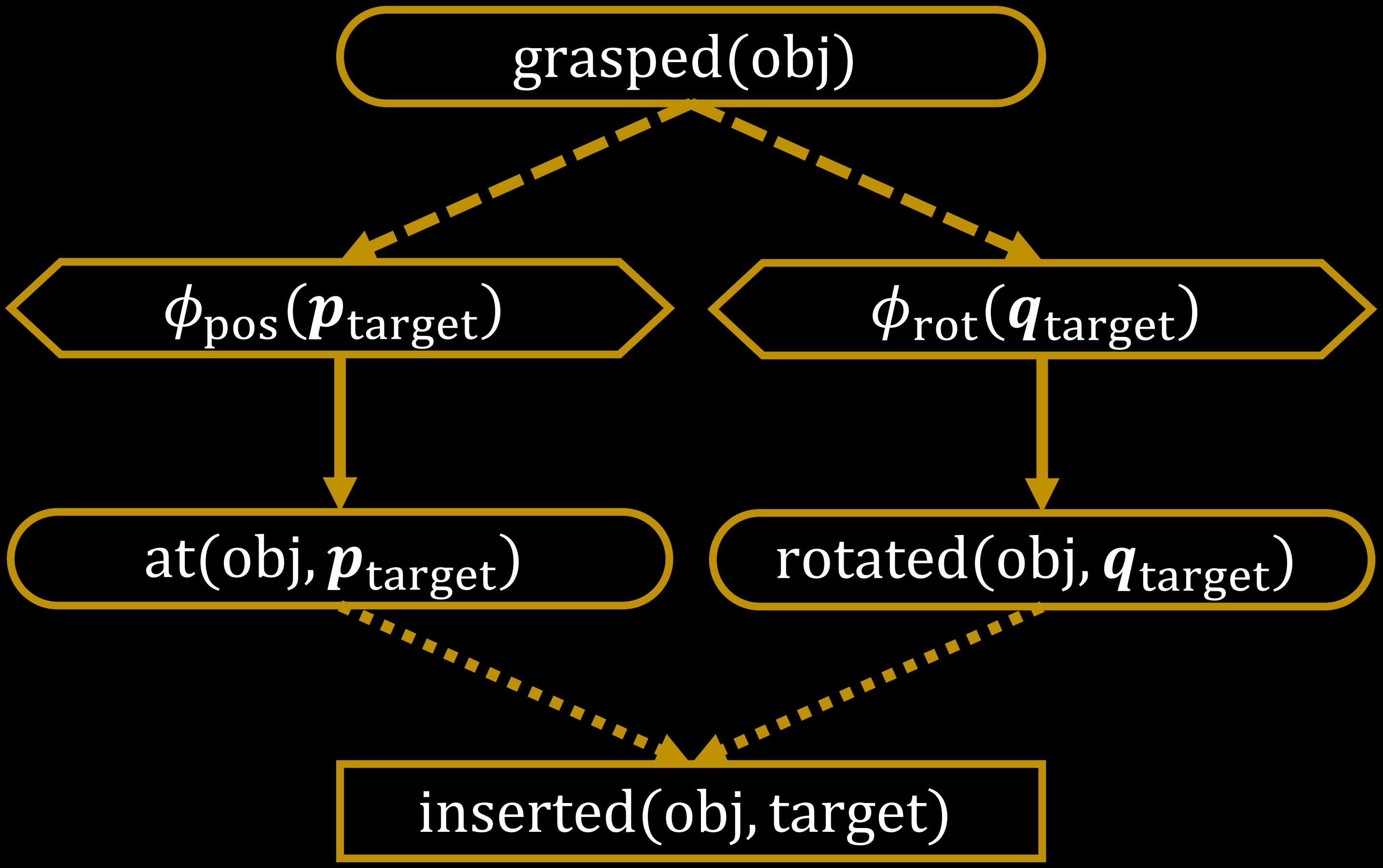


Composed Causal Graphs

The robot will autonomously estimate the state-action utility of executing each possible composition to achieve the intended effects in the assembly task.

$$\phi_{\mathrm{pos}} \lhd \phi_{\mathrm{rot}}$$

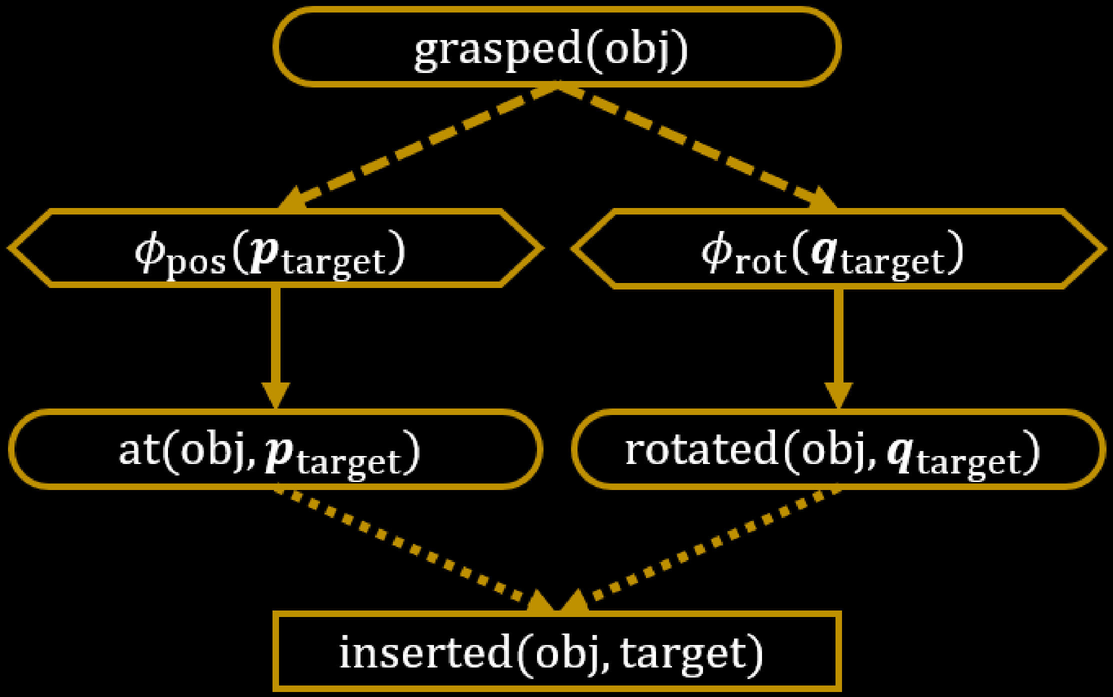
$$\underbrace{\phi_{\mathrm{rot}}}_{\text{left-hand side: lower-priority}} \lhd \underbrace{\phi_{\mathrm{pos}}}_{\text{right-hand side: higher-priority}}$$



Furniture Part Connection Policy

To estimate how well each multi-objective action a will achieve its composed effects, the robot will perform N Monte Carlo simulations to estimate the state-action utility of possible compositions

$$\hat{Q}(s,a) = \frac{1}{N}\sum_{n=1}^{N}\hat{Q}_n(s,a)$$

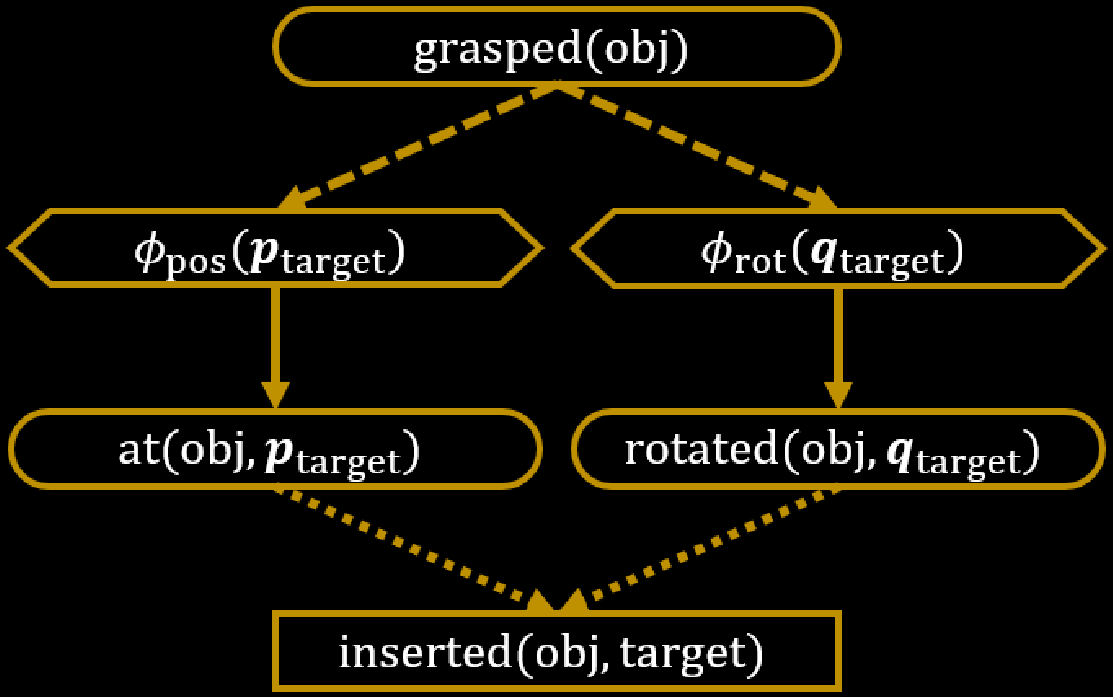


Furniture Part Connection Policy

During task execution, the robot will choose the behavior with the maximum predicted state-action utility based on the Monte Carlo simulations:

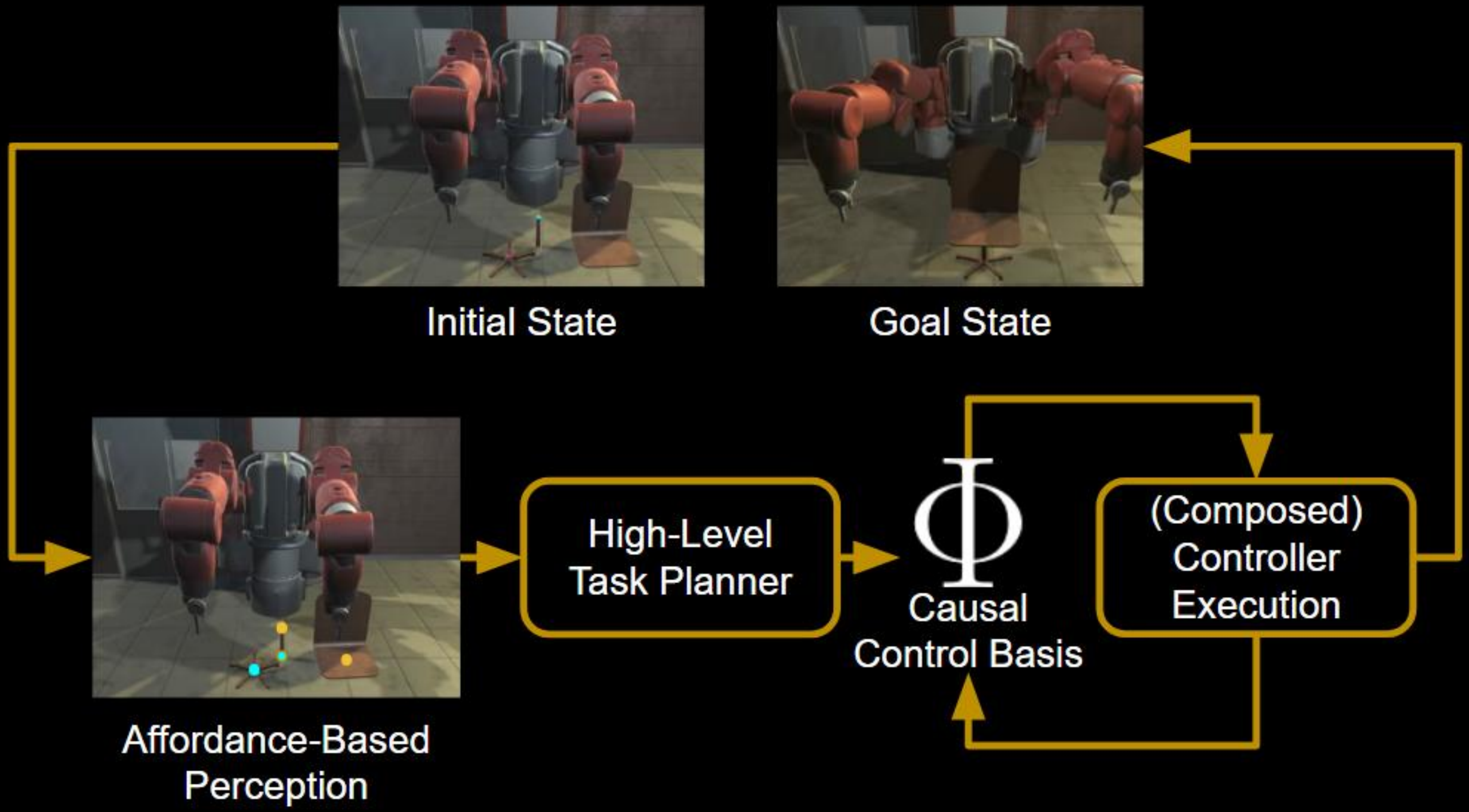
$$\hat{Q}(s,a) = \frac{1}{N}\sum_{n=1}^{N}\hat{Q}_n(s,a)$$

$$a = \pi(s) = \arg\max_{a}\left(\hat{Q}(s,a)\right)$$



[22] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach, Fourth Edition*, Pearson Education, 2020.
[23] M. L. Litman, T. L. Dean, and L. P. Kaelbling, "On the Complexity of Solving Markov Decision Problems," *arXiv preprint arXiv:1302.4971*, 2013.

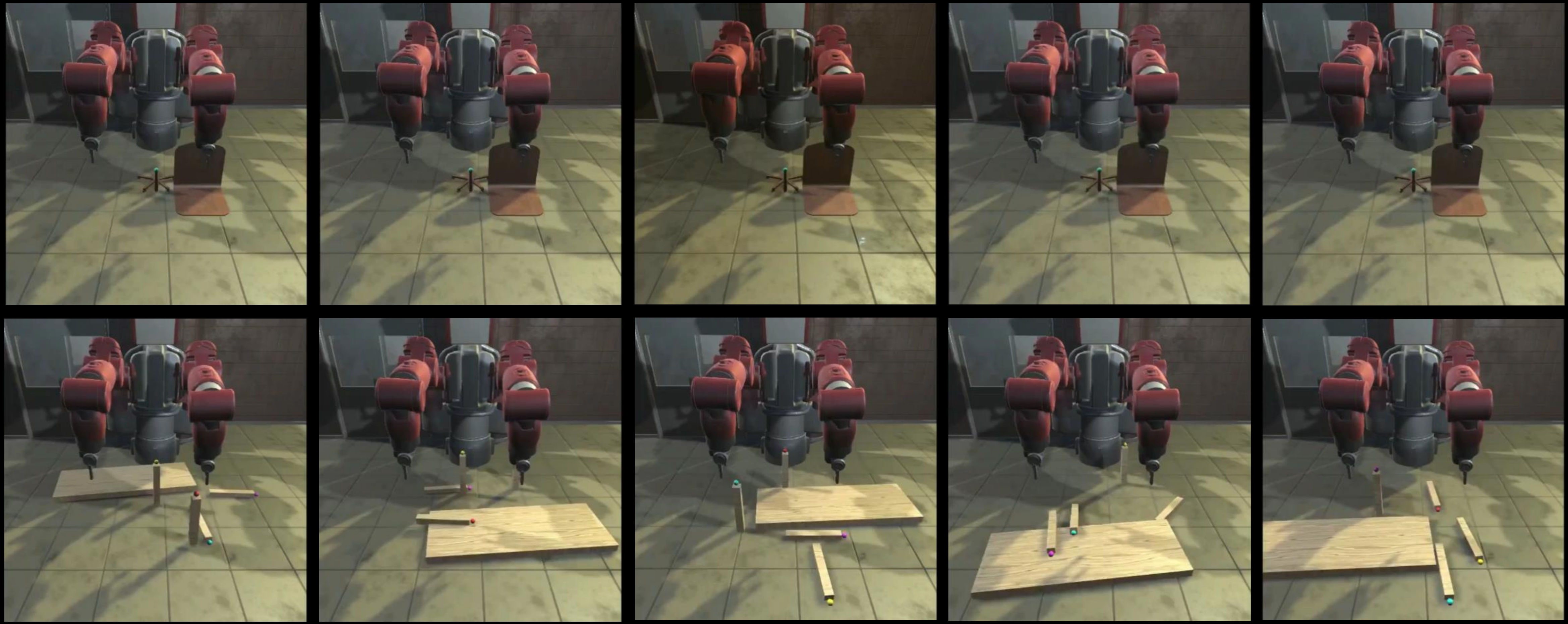
Causal Control Basis for Furniture Assembly



We use an off-the-shelf high-level task planner to sequence high-level actions. The causal control basis describes how to sequence low-level controllers and how to autonomously compose the multi-objective connection actions.

[24] Y. Lee, E. S. Hu, and J. J. Lim, "IKEA Furniture Assembly Environment for Long-Horizon Complex Manipulation Tasks," *IEEE ICRA*, 2021.

Assembly Experiments



The causal control basis selected composition $\phi_{\text{pos}} \lhd \phi_{\text{rot}}$ for composed effect inserted and composition $\phi_{\text{rot}} \lhd \phi_{\text{screw}} \lhd \phi_{\text{pos}}$ for composed effect screwed-in.

Furniture Assembly Results

The multi-objective actions the causal control basis predicted to achieve the composed effects enabled the robot to perform furniture assembly tasks with reasonable success.

Connection Action	Successful Connections	Connection Attempts	Success Rate
Insert	20	28	0.714
Screw	40	42	0.952
TOTAL	**60**	**70**	**0.857**

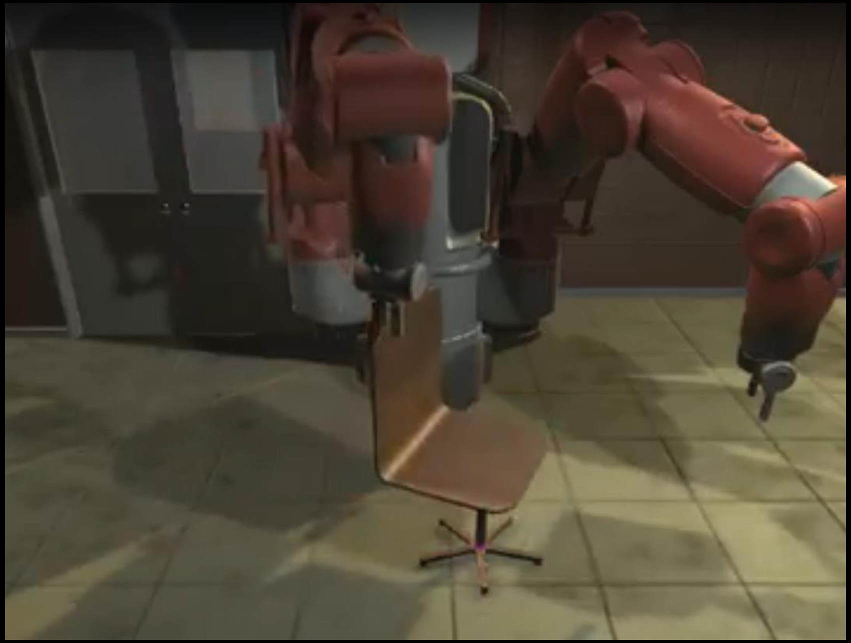
The results provide evidence that the causal control basis effectively captures causal information relevant for autonomously composing controllers for complex behaviors.

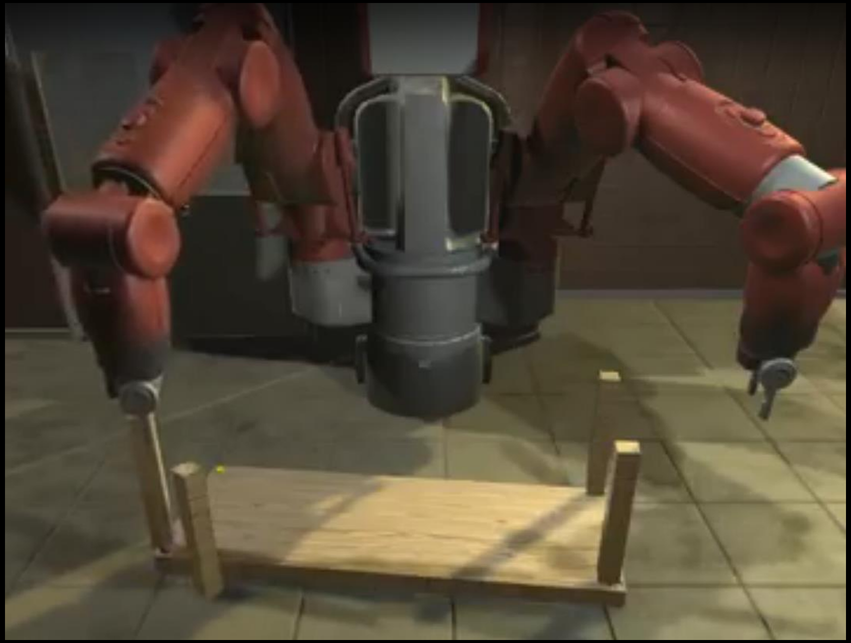
[25] E. Sheetz *et al.*, "Composable Causality in Semantic Robot Programming," *IEEE ICRA*, 2022.

Future Work for Composed Causality

Future work beyond the scope of the dissertation includes:

- Evaluation in real-world assembly tasks
- Extending to tasks that require whole-body manipulation
- More detailed failure analysis to determine patterns in performance

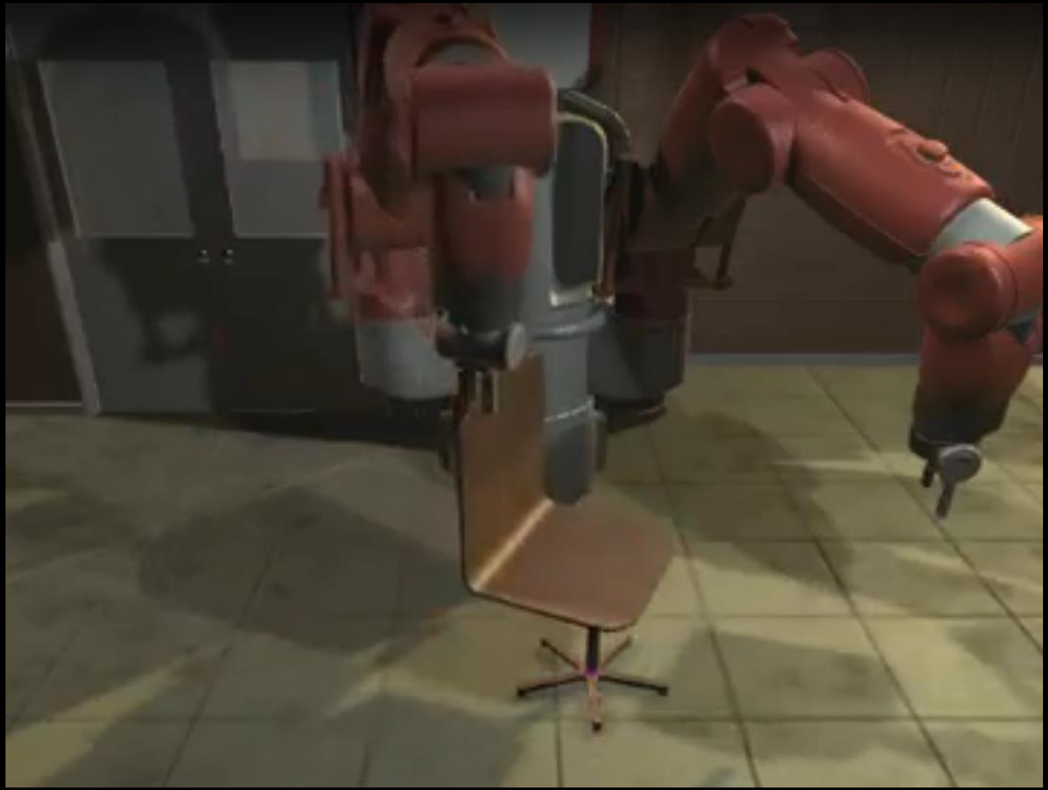




Planning Complex Actions

We demonstrate that the causal control basis effectively provides causal information for autonomous controller composition (ICRA 2022).

The causal control basis uses explainable cause-effect relationships to minimize the expert knowledge required to perform complex tasks.



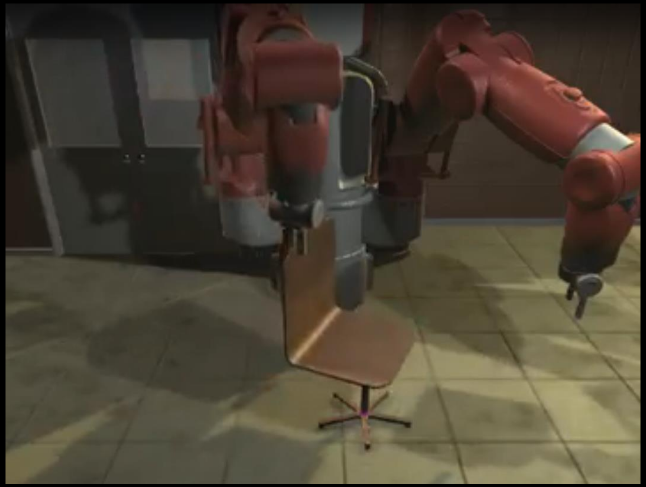
[25] E. Sheetz *et al.*, "Composable Causality in Semantic Robot Programming," *IEEE ICRA*, 2022.

Dissertation Contributions

Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

Safety reasoning on human-robot teams
(UR 2025, Under Review)

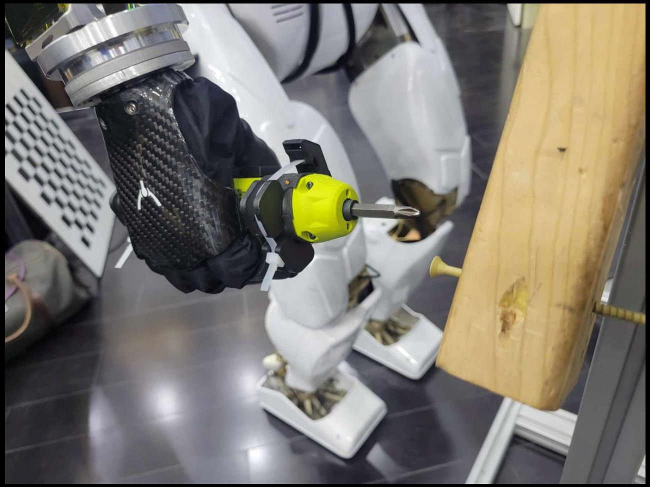


Dissertation Contributions

Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

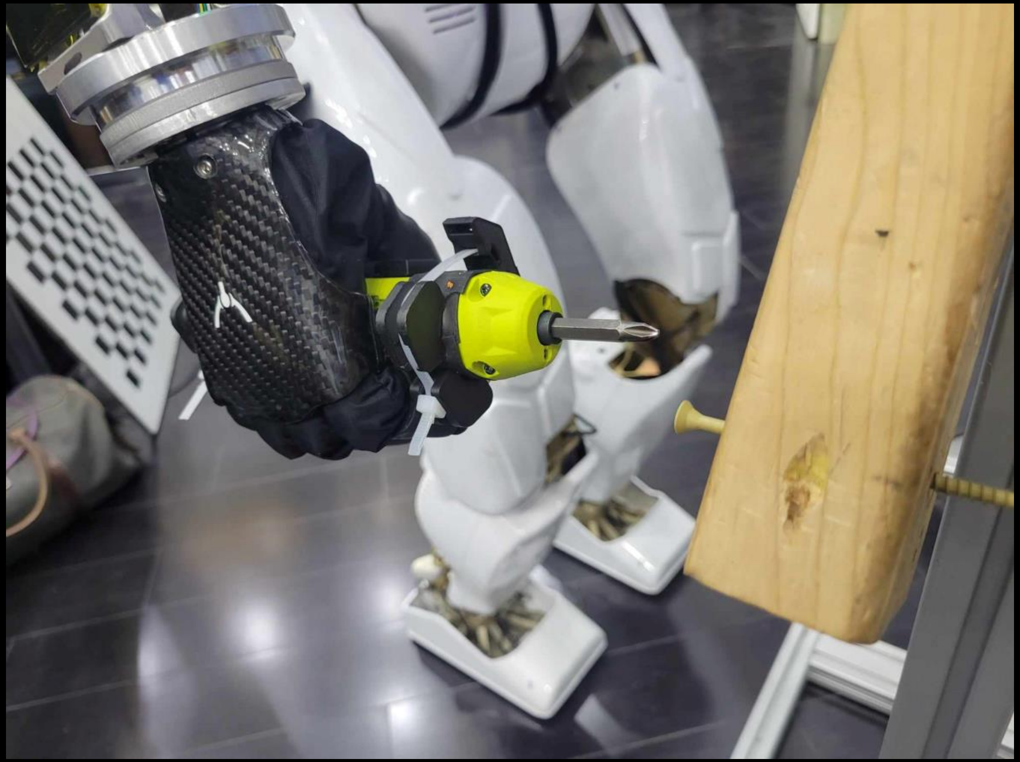
Safety reasoning on human-robot teams
(UR 2025, Under Review)



Reliable and Explainable Actions: Grasp Reflex Model

Reliable and Explainable Behaviors

To promote understanding on human-robot teams, we need the complex actions in tool-use and assembly tasks to be explainable.



[26] Y. Zhang *et al.*, "Plan Explicability for Robot Task Planning," *RSS Workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics*, 2016.

Robot Grasping

Robot grasping achieves contacts and forces to restrain objects for manipulation.

Multi-fingered end-effectors provide abundant sensor signals and degrees-of-freedom for performing dexterous manipulations.

[27] A. Bicchi and V. Kumar, "Robotic Grasping and Contact: A Review," *IEEE ICRA*, 2000.
[28] T. Mouri, H. Kawasaki, and S. Ito, "Unknown Object Grasping Strategy Imitating Human Grasping Reflex for Anthropomorphic Robot Hand," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, 2007.

Challenge: Explainability in Action Models

Data-driven approaches for learning and modeling actions show significant promise and great performance.

But neural network and deep learning approaches tend to be black-box models that lead to poor understanding on human-robot teams in safety-critical domains.

[26] Y. Zhang *et al.*, "Plan Explicability for Robot Task Planning," *RSS Workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics*, 2016.
[29] NASA, "NASA Risk Management Handbook," [Online], 2011.
[30] NASA, "NASA Safety Culture Handbook," [Online], 2015.

Insight: Human Grasp Reflex

We take inspiration from the human grasp reflex, specifically the involuntary newborn palmar reflex.

Similar reflex control approaches map sensory data to learned patterns of response.

We aim to learn a reflex model for grasping that reduces knowledge engineering while remaining explainable.

[31] A. Anekar and B. Bordoni, "Palmar Grasp Reflex," *StatPearls Publishing*, 2012.
[32] Y. Futagi, Y. Toribe, and Y. Sazuki, "The Grasp Reflex and Moro Reflex in Infants: Hierarchy of Primitive Reflex Responses," *International Journal of Pediatrics*, 2012.
[33] G. Bekey and R. Tomovic, "Robot Control by Reflex Actions," *IEEE ICRA*, 1986.

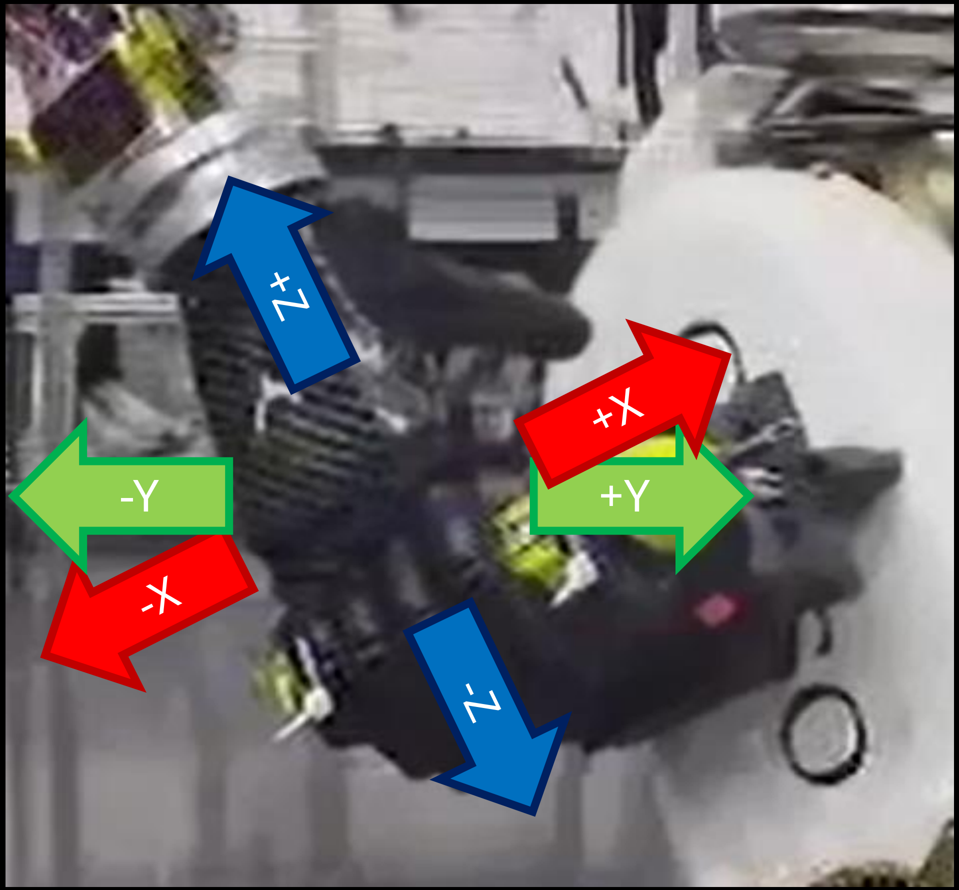
Grasp Reflex Modeling

We propose a simple, explainable grasp reflex model that allows the robot to adjust its grasp on a tool until it is secure enough for a tool manipulation task.

Grasp Reflex Model

The grasp reflex model uses a simple logistic regression model to map continuous end-effector joint states to discrete symbolic adjustment states.

The known symbolic adjustment states are prerequisite states for adjustment actions that allow the robot to improve its grasp on the tool.



[34] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems*, 2009.

Grasping Novel Tools

training drill

testing tools

After learning the grasp reflex model on a training tool, we performed experiments to test how well the learned reflex generalized to grasping novel test tools.

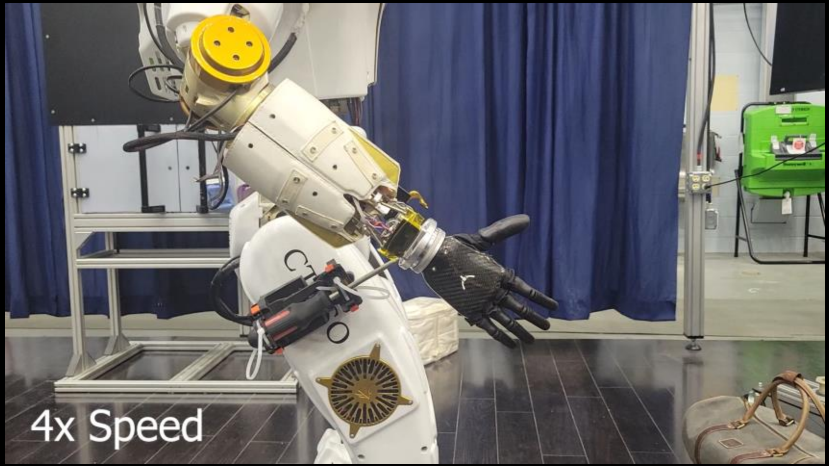
Grasping Novel Tools

We provide one reference joint configuration for each test tool as an example of a secure grasp.

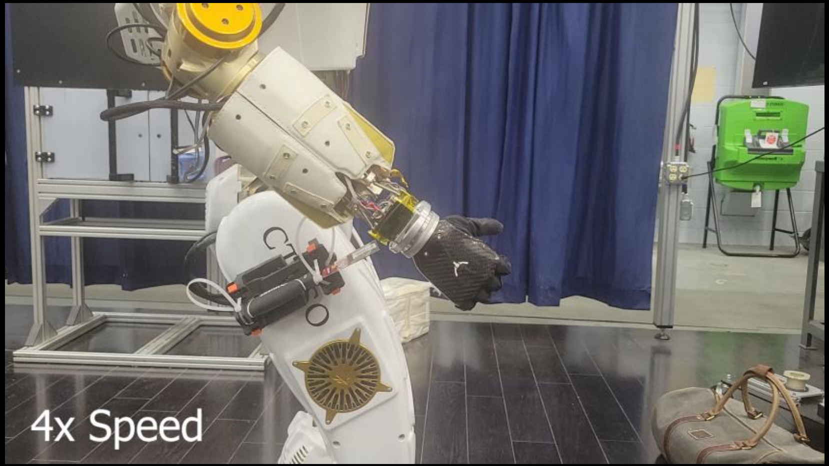
The robot repeatedly attempts grasps and adjusts its grasp until the grasp reflex model predicts the grasp is secure.

One-Shot Tactile Servoing on Novel Tools

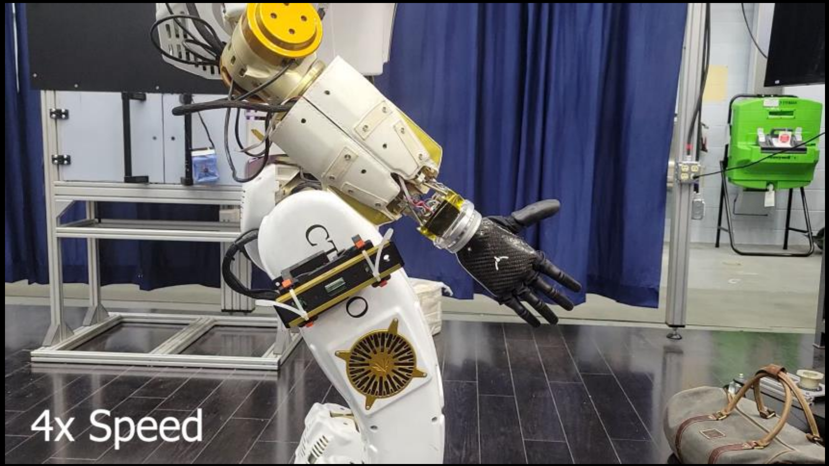
Screwdriver SUCCESS!



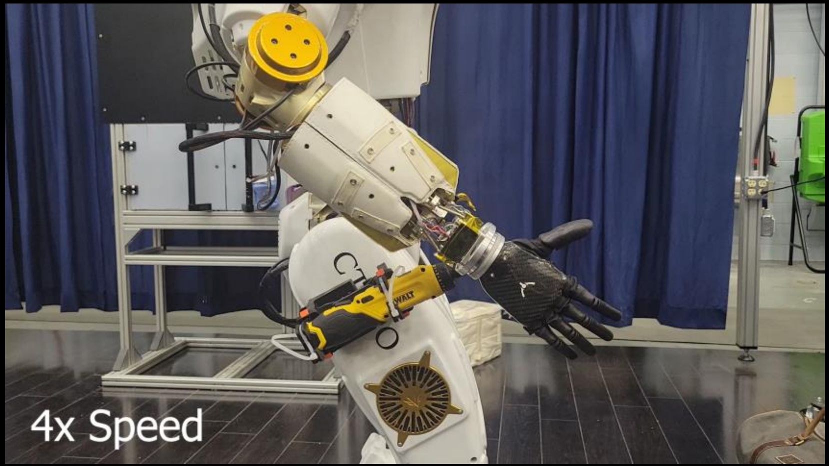
Paint Scraper FAILED



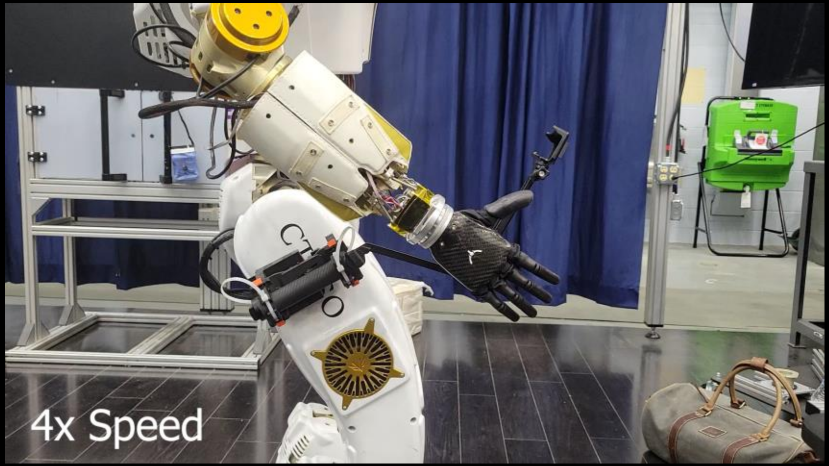
Level FAILED



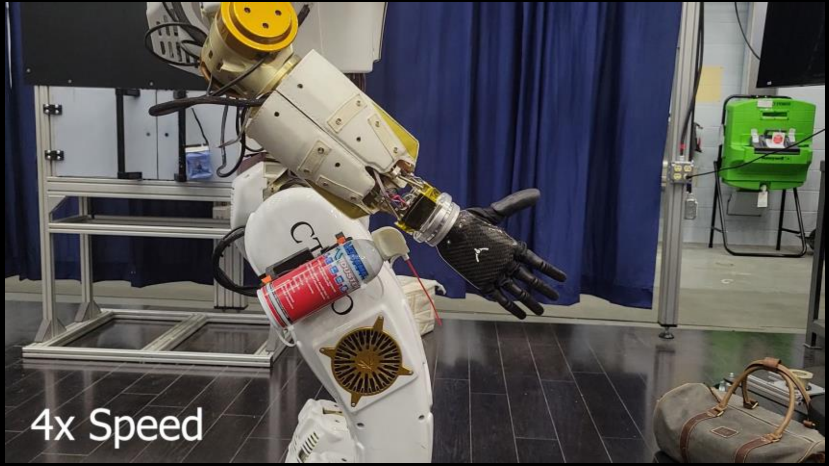
Gyroscopic Drill SUCCESS!



Selfie Stick SUCCESS!



Compressed Air Can SUCCESS!



Grasp Reflex Model Results

We evaluated in-hand (the robot did not drop the tool) and manipulation (secure enough for subsequent tool-use tasks) grasps.

Tool	Practical for End-Effector	In-Hand Grasp Success Rate	Manipulation Grasp Success Rate
Drill	Yes	1.00	1.00
Screwdriver	Yes	1.00	0.83
Paint Scraper	Yes	1.00	0.67
Level	Yes	0.83	0.67
Gyroscopic Drill	Yes	1.00	0.50
Selfie Stick	**No**	1.00	0.33
Compressed Air Can	**No**	1.00	0.17
CUMULATIVE	**-**	**0.98**	**0.60**
PRACTICAL CUMULATIVE	**-**	**0.97**	**0.73**

[35] E. Sheetz *et al.*, "Multi-Fingered End-Effector Grasp Reflex Modeling for One-Shot Tactile Servoing in Tool Manipulation Tasks," *IEEE IROS*, 2024.

Grasp Reflex Model Results

The results show promise for a simple, inherently explainable action reflex. However, we may need to model distributions of tool features (size, graspable surface area, weight distribution).

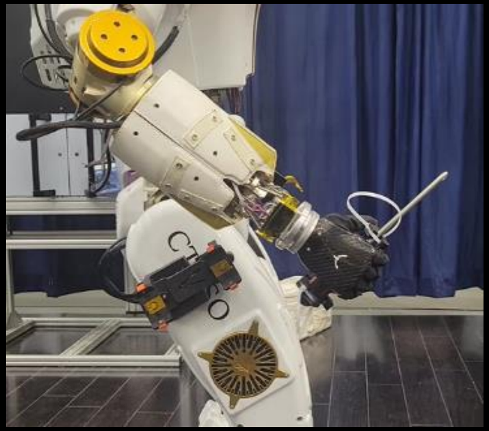
Tool	Practical for End-Effector	In-Hand Grasp Success Rate	Manipulation Grasp Success Rate
Drill	Yes	1.00	1.00
Screwdriver	Yes	1.00	0.83
Paint Scraper	Yes	1.00	0.67
Level	Yes	0.83	0.67
Gyroscopic Drill	Yes	1.00	0.50
Selfie Stick	**No**	1.00	0.33
Compressed Air Can	**No**	1.00	0.17
CUMULATIVE	**-**	**0.98**	**0.60**
PRACTICAL CUMULATIVE	**-**	**0.97**	**0.73**

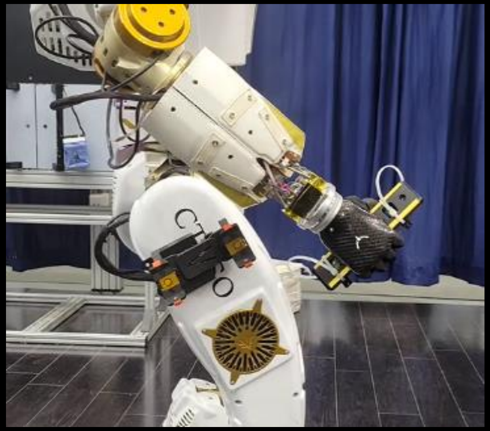
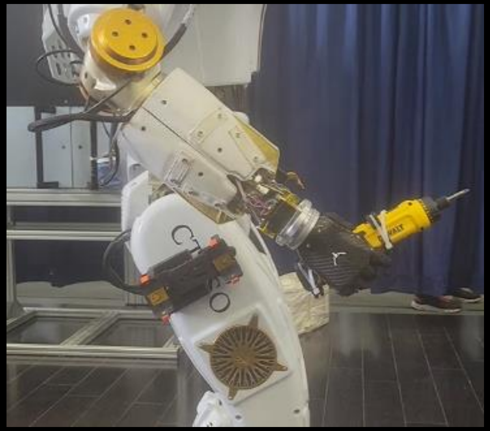
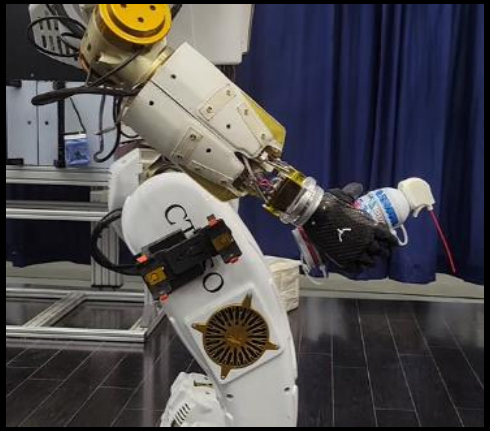
[35] E. Sheetz *et al.*, "Multi-Fingered End-Effector Grasp Reflex Modeling for One-Shot Tactile Servoing in Tool Manipulation Tasks," *IEEE IROS*, 2024.

Future Work for Grasp Reflex Modeling

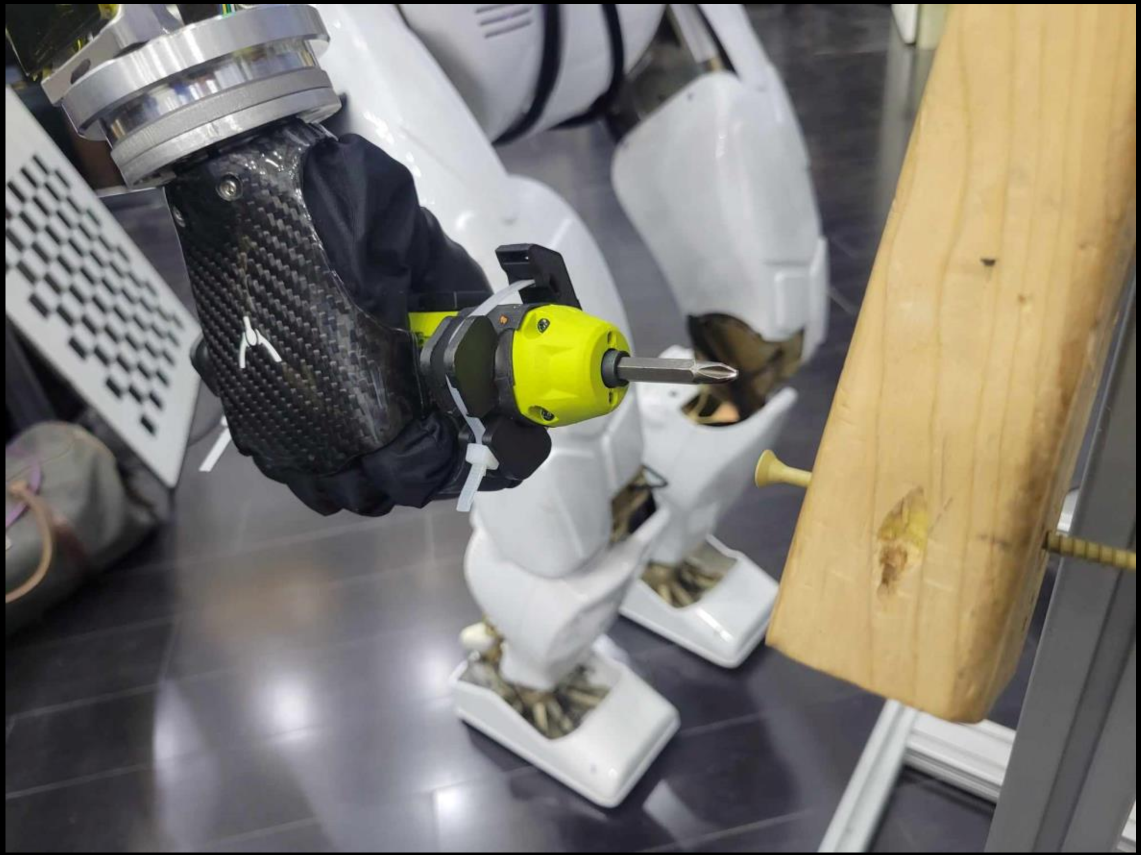
Future work beyond the scope of the dissertation includes:

- Training over a set of representative tools and features
- Learning different types of grasps (precision, trigger) and adjustment actions
- Autonomous exploration or “play” to learn about different tools



Explainable Actions



We demonstrate the promise of a simple, inherently explainable grasp reflex model for achieving reliable performance and generalizable behaviors (IROS 2024).

The grasp reflex model uses explainable symbolic adjustments to promote understandable action execution on human-robot teams.

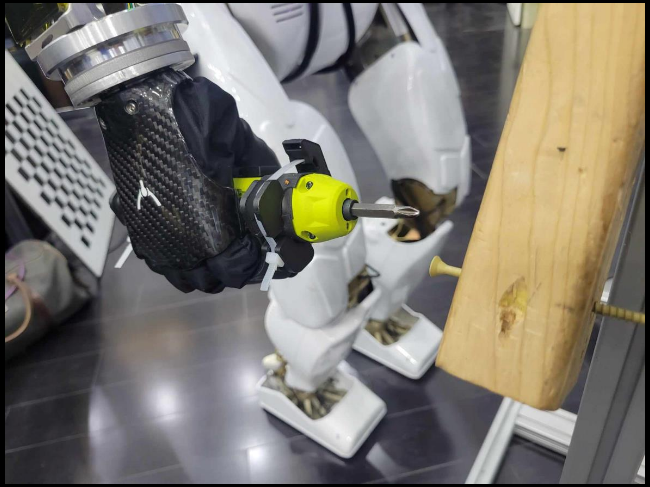
[35] E. Sheetz *et al.*, "Multi-Fingered End-Effector Grasp Reflex Modeling for One-Shot Tactile Servoing in Tool Manipulation Tasks," *IEEE IROS*, 2024.

Dissertation Contributions

Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

Safety reasoning on human-robot teams
(UR 2025, Under Review)

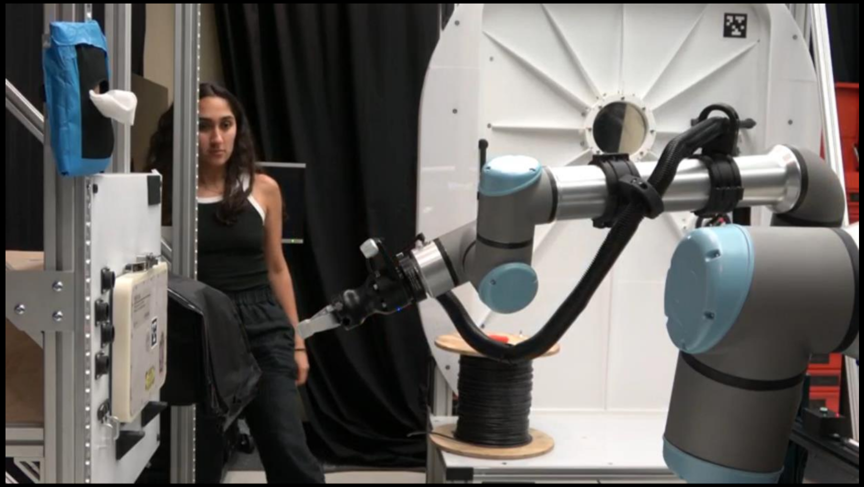


Dissertation Contributions

Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

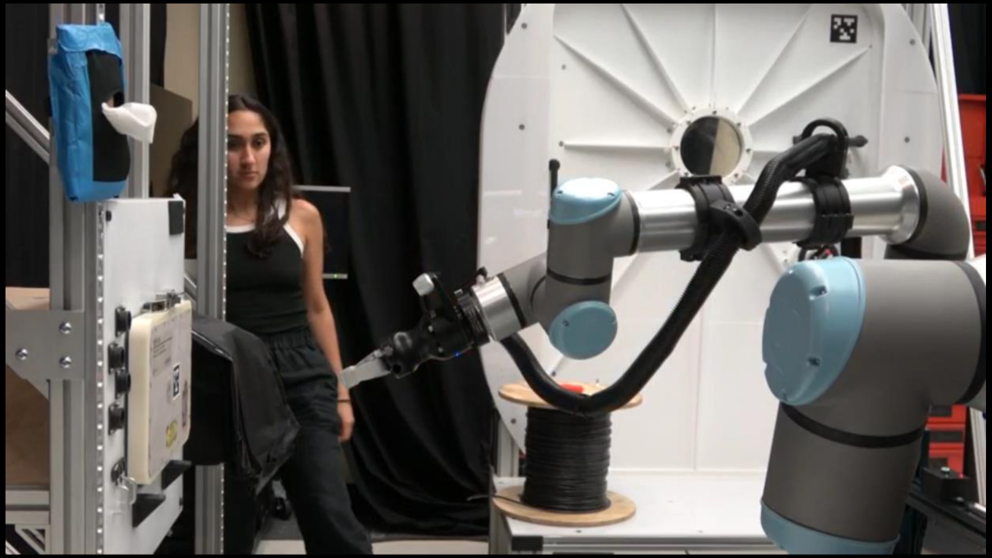
Safety reasoning on human-robot teams
(UR 2025, Under Review)



Safety Reasoning:
Human-Robot Red Teaming

Safety and Trust

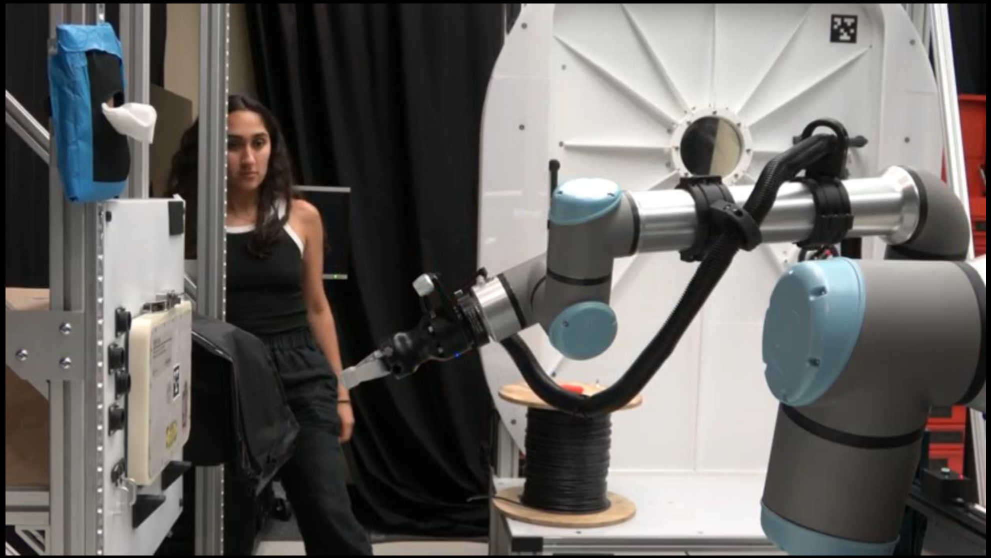
For robots to be effective performing cooperative tasks in safety-critical domains, we expect robots to earn trust on human-robot teams.



[5] M. Vasic and A. Billard, “Safety Issues in Human-Robot Interactions,” *IEEE ICRA*, 2013.
[6] Y. Zhang *et al.*, “DANLI: Deliberative Agent for Following Natural Language Instructions,” *arXiv preprint arXiv:2210.12485*, 2022.
[9] B. Kuipers, “Trust and Cooperation,” *Frontiers in Robotics and AI*, 2022.
[36] BS Dhillon, ARM Fashandi, and KL Liu, “Robot Systems Reliability and Safety: A Review,” *Journal of Quality in Maintenance Engineering*, 2002.

Challenge: Safety Reasoning

Despite the consensus on the importance of robot safety, much research overtrusts the robot's capabilities and/or the human operators to guarantee safe operations.



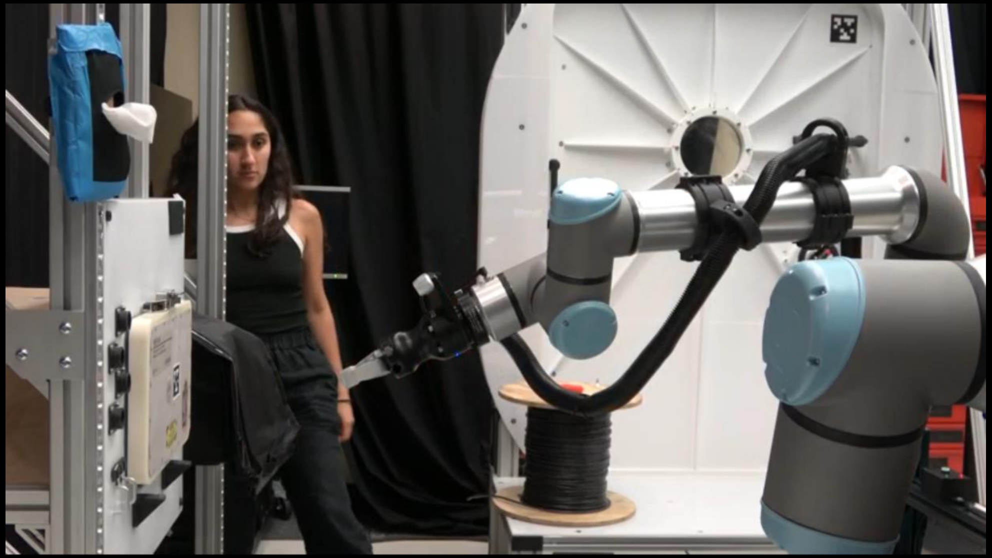
[5] M. Vasic and A. Billard, "Safety Issues in Human-Robot Interactions," *IEEE ICRA*, 2013.
[6] Y. Zhang *et al.*, "DANLI: Deliberative Agent for Following Natural Language Instructions," *arXiv preprint arXiv:2210.12485*, 2022.
[9] B. Kuipers, "Trust and Cooperation," *Frontiers in Robotics and AI*, 2022.
[36] BS Dhillon, ARM Fashandi, and KL Liu, "Robot Systems Reliability and Safety: A Review," *Journal of Quality in Maintenance Engineering*, 2002.

Insight: Red Teaming

Robots use models to simplify reasoning in an unboundedly complex world.

While simplifying models are useful, disastrous outcomes occur when a critical factor is left out of the model.

Red teaming considers adversarial perspectives to improve decision making.



[37] B. Kuipers, "AI and Society: Ethics, Trust, and Cooperation, *Communications of the ACM*, 2023.
[38] A. Yang *et al.*, "Characterizing Warfare in Red Teaming," *IEEE Systems, Man, and Cybernetics*, 2006.
[39] D. F. Longbine, "Red Teaming: Past and Present," *School of Advanced Military Studies, Army Command and General Staff College*, 2008.
[40] M. Zenko, *Red Team: How to Succeed by Thinking Like the Enemy*, Basic Books, 2015.

Computational Red Teams

Computational red teams (CRTs) are teams of computational agents that automate the adversary red team trying to thwart the blue team's objective. The CRT helps improve decision making on the blue team.

[41] D. Ganguli *et al.*, "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," *arXiv preprint arXiv:2209.07858*, 2022.
[42] E. Perez *et al.*, "Red Teaming Language Models with Language Models," *arXiv preprint arXiv:2202.03286*, 2022.

CRTs for Safety-Critical Tasks

Our preliminary experiments with the current state-of-the-art computational agents indicate that fully automated CRTs may not effectively update modeled knowledge. Furthermore, research suggests humans are necessary for evaluative moral and ethical judgments.

Blue Team
(ChatGPT)

Red Team
(English-like chatbot)

[43] T. B. Sheridan, "Human-Robot Interaction: Status and Challenges," *Human Factors*, 2016.
[44] B. Kuipers, "How Can We Trust a Robot?," *Communications of the ACM*, 2018.
[45] B. Kuipers, "Perspectives on Ethics of AI," *The Oxford Handbook of Ethics of AI*, Oxford University Press, 2020.

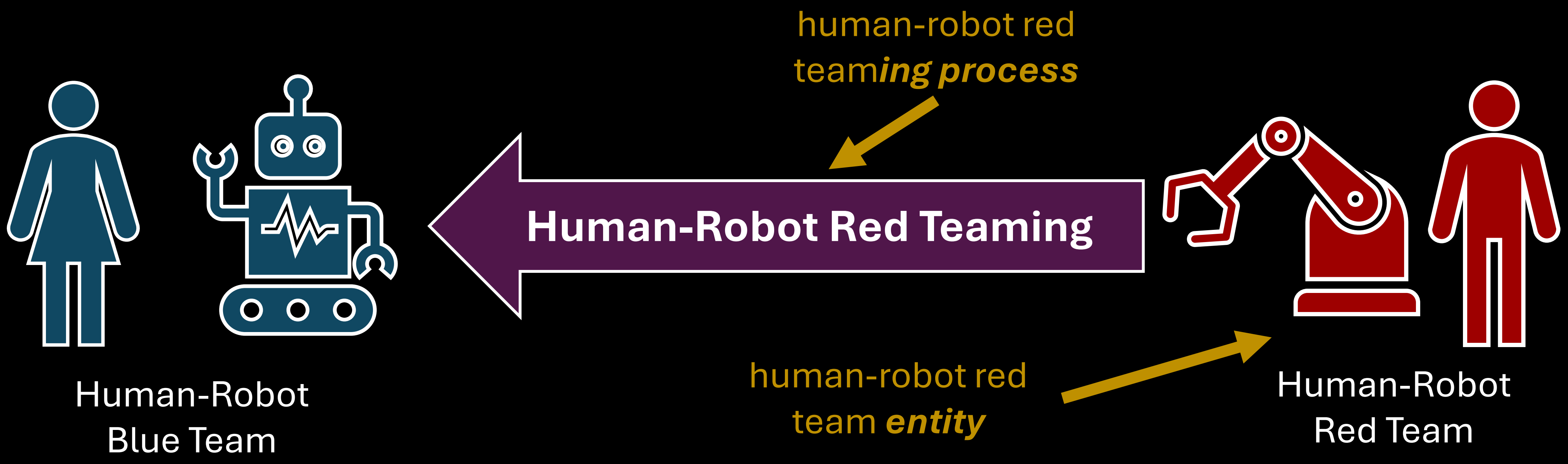
Human-Robot Red Teaming Approach

Human-Robot Blue Team

Human-Robot Red Team

To overcome challenges faced by computational red teams, we propose a human-robot red team (HRRT) to allow human and robot agents to collaboratively analyze safety in shared autonomy tasks.

Human-Robot Red Teaming Approach



The HRRT (as a subset of CRT) does not act as an adversary thwarting the blue team's objectives, but rather a challenger to the human-robot blue team's modeled knowledge, expectations, assumptions, and contingency plans.

Levels of Computational Red Teaming

Computational red teams (CRTs) are categorized according to their level of reasoning:

- **CRT0**: Simple decision-making agents do not evolve.
- **CRT1**: Agents learn and adapt.
- **CRT2**: Teams of agents learn and adapt together.
- **CRT3**: Teams evolve within a dynamic environment.
- **CRT4**: Teams reflect and unlearn their biases to learn better approaches.

[46] H. Abbass *et al.*, “Computational Red Teaming: Past, Present, and Future,” *IEEE Computational Intelligence Magazine*, 2011.

Levels of Red Teaming

Computational red teams (CRTs) are categorized according to their level of reasoning:

- **CRT0**: Simple decision-making agents do not evolve.
- **CRT1**: Agents learn and adapt.
- **CRT2**: Teams of agents learn and adapt together.
- **CRT3**: Teams evolve within a dynamic environment.
- **CRT4**: Teams reflect and unlearn their biases to learn better approaches.

We suggest that human-robot red teaming will similarly benefit from multiple levels of capability to characterize responsibilities.

[46] H. Abbass *et al.*, "Computational Red Teaming: Past, Present, and Future," *IEEE Computational Intelligence Magazine*, 2011.

HRRTs as Subsets of CRTs

Computational red teams (CRTs) are categorized according to their level of reasoning:

- **CRT0**: Simple decision-making agents do not evolve.
- **CRT1**: Agents learn and adapt.
- **CRT2**: Teams of agents learn and adapt together.
- **CRT3**: Teams evolve within a dynamic environment.
- **CRT4**: Teams reflect and unlearn their biases to learn better approaches.

We observe that some of the CRT levels focus on teams of agents and propose comparable levels to human-robot red teaming, where HRRTs are specific subsets of CRTs where computational agents work on teams alongside humans.

[46] H. Abbass *et al.*, "Computational Red Teaming: Past, Present, and Future," *IEEE Computational Intelligence Magazine*, 2011.

Levels of Human-Robot Red Teaming

Human-robot red teams (HRRTs) are categorized according to their level of reasoning:

- **HRRT2**: Teams of human and robot agents learn and adapt together by enumerating possibilities given their knowledge of the environment.
- **HRRT3**: Teams of human and robot agents evolve within a dynamic environment by challenging assumptions implicit in their modeled knowledge.
- **HRRT4**: Teams of human and robot agents reflect together and improve modeled knowledge to address “unknown unknowns.”

[9] B. Kuipers, “Trust and Cooperation,” *Frontiers in Robotics and AI*, 2022.

Overview of HRRT Levels and Iterations

Current Model
$\boldsymbol{M = (S, A)}$

Model M is set of symbolic states S and actions A that describe the robot's reasoning in an environment

A complete model M^* of an unboundedly complex world is intractable, so the robot reasons over simplified model $M \subset M^*$

We need to ensure model M allows the team to adequately reason about safety, so we analyze what may be left out of M to create updated model M'

Overview of HRRT Levels and Iterations

Current Model
$\boldsymbol{M = (S, A)}$

Level HRRT2: analyze possible states in M

HRRT2 identifies state transitions, however unlikely, and prompts the team to reflect on the validity of these possibilities and if there are expected possibilities not reflected by the current model

$$\mathcal{H}_2(M) = \{(s, a, s') | s, s' \in S(M), a \in A(M), \\ \texttt{actionable}(a, s), \texttt{effect}(a, s')\}$$

Overview of HRRT Levels and Iterations

Current Model
$\boldsymbol{M = (S, A)}$

Level HRRT2: analyze possible states in M

Level HRRT3: analyze action assumptions in M

HRRT3 identifies implicit pre- and post-condition assumptions, prompting the team to consider contingency plans

$$\mathcal{H}_3(M) = (\Omega_{\text{pre}}, \Omega_{\text{post}})$$

$$\Omega_{\text{pre}} = \{\omega_{\text{pre}} = (s, a) | s \in S(M), a \in A(M), \texttt{precond}(s, a)\}$$
$$\Omega_{\text{post}} = \{\omega_{\text{post}} = (a, s) | s \in S(M), a \in A(M), \texttt{postcond}(a, s)\}$$

Overview of HRRT Levels and Iterations

Current Model
$\boldsymbol{M = (S, A)}$

Level HRRT2:
analyze possible states in M

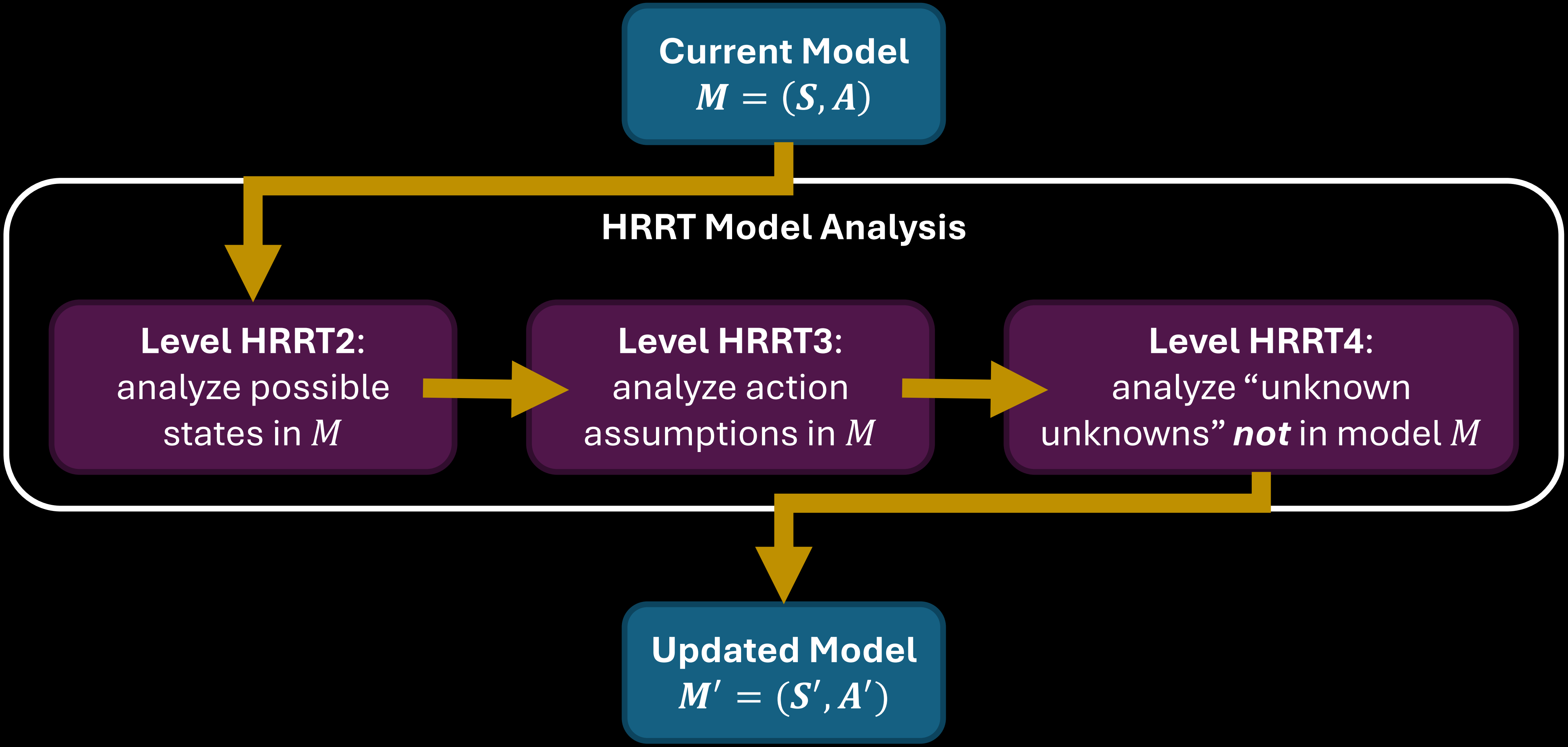
Level HRRT3:
analyze action assumptions in M

Level HRRT4:
analyze “unknown unknowns” ***not*** in model M

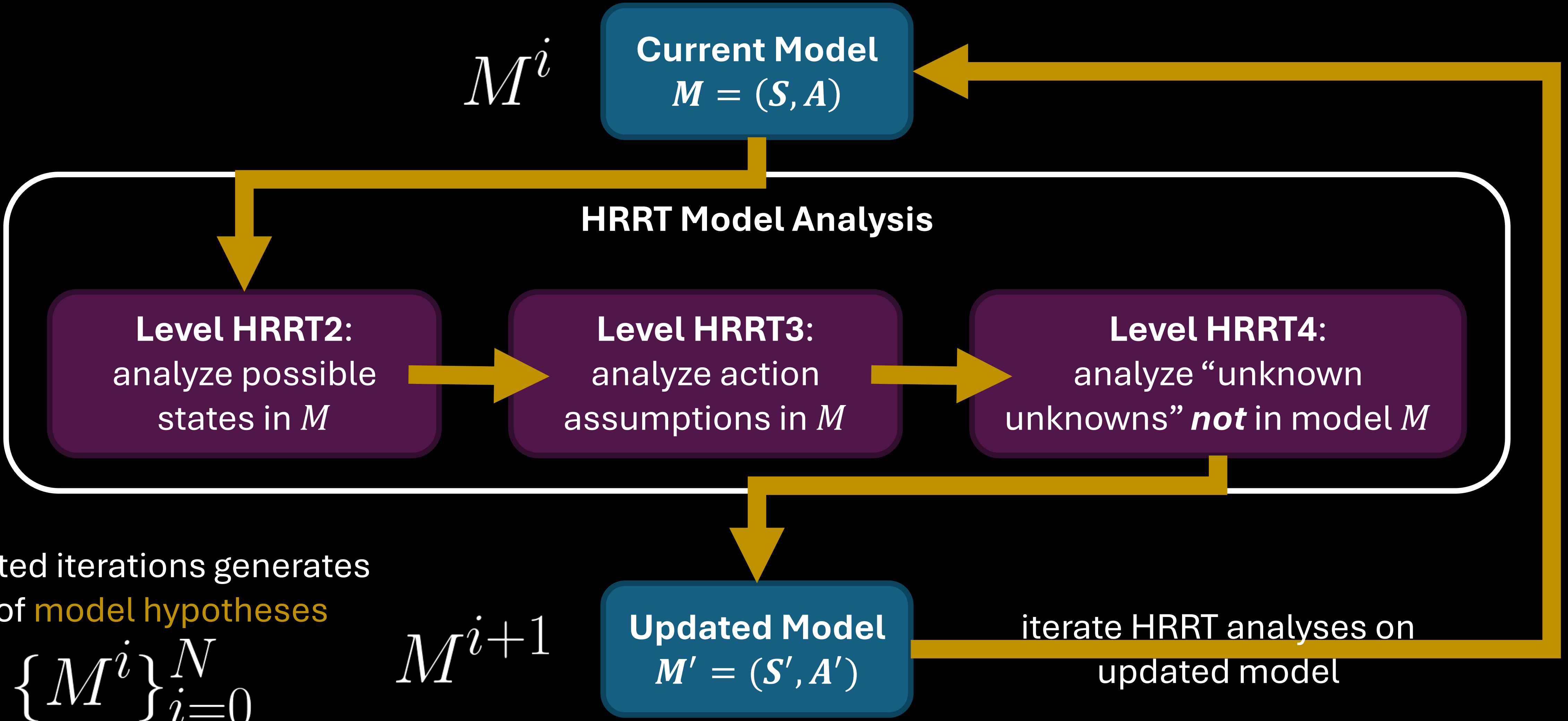
$$\mathcal{H}_4(M, \mathcal{H}_2(M), \mathcal{H}_3(M), \Sigma) = M'$$

HRRT4 uses dialogue prompts in Σ to prompt deeper reflections on general safety, domain-specific questions, and “unknown unknowns”

Overview of HRRT Levels and Iterations



Overview of HRRT Levels and Iterations



Repeated iterations generates set of model hypotheses

$\{M^i\}_{i=0}^N$

HRRT Experiments Overview

Human-Robot
Blue Team
(ChatGPT,
direction from researcher)

Human-Robot
Red Team
(automated methods,
dialogue tree chatbot)

Given basic information about the domain M^0, the red robot agents query the human-robot blue team to update the team's modeled knowledge. This process assumes the blue team (specifically human agents) have some perspective or insight about the domain.

[47] OpenAI, ChatGPT, [Online], 2025.

HRRT Experiments Overview

Human-Robot
Blue Team
(ChatGPT,
direction from researcher)

Human-Robot
Red Team
(automated methods,
dialogue tree chatbot)

Through simple English-like interactions, the human-robot team explores safety in different problem domains.

We iterate through the human-robot red teaming exercise, saving the model hypothesis for each iteration.

HRRT Experiments Overview

Human-Robot
Blue Team
(ChatGPT,
direction from researcher)

Human-Robot
Red Team
(automated methods,
dialogue tree chatbot)

All new symbols to the model hypotheses are proposed by the blue computational ChatGPT agent, as a result of prompts from the red computational chatbot agent.

Example Interaction

Consider a team of robots conducting science experiments on the surface of Mars, communicating with ground control on Earth over time delay.

[48] E. Sheetz, “Tool-Use Robot Manipulation Tasks for Cooperative and Explainable Operations in Safety-Critical Domains,” PhD Dissertation, University of Michigan, 2025.

Example Interaction

Consider a team of robots conducting science experiments on the surface of Mars, communicating with ground control on Earth over time delay.

HRRT2: Analysis of Possibilities

Red Computational Agent Prompt

The model indicates that if `mission_interrupted`, then `robot_available` should not be possible.

Human-Robot Blue Team Response

Correct, a mission interruption means the robot should not be available to start new tasks.

[48] E. Sheetz, "Tool-Use Robot Manipulation Tasks for Cooperative and Explainable Operations in Safety-Critical Domains," PhD Dissertation, University of Michigan, 2025.

Example Interaction

Consider a team of robots conducting science experiments on the surface of Mars, communicating with ground control on Earth over time delay.

HRRT3: Analysis of Assumptions

Red Computational Agent Prompt

The model includes action `activate_redundant_comms`, which should require pre-condition `comm_blackout`.

Human-Robot Blue Team Response

Incorrect, redundancy will be valuable in this domain, even without failures or blackouts.

[48] E. Sheetz, "Tool-Use Robot Manipulation Tasks for Cooperative and Explainable Operations in Safety-Critical Domains," PhD Dissertation, University of Michigan, 2025.

Example Interaction

Consider a team of robots conducting science experiments on the surface of Mars, communicating with ground control on Earth over time delay.

HRRT4: Analysis of "Unknown Unknowns"

Red Computational Agent Prompt	Human-Robot Blue Team Response
Are there additional tasks the robot should be taking on?	Yes, microbial contamination detection and solar panel cleaning.
What catastrophic failures could occur in this domain?	Major data loss or structural damage due to environment.

Team updates the model based on these reflections.

[48] E. Sheetz, "Tool-Use Robot Manipulation Tasks for Cooperative and Explainable Operations in Safety-Critical Domains," PhD Dissertation, University of Michigan, 2025.

Example Interaction

Consider a team of robots conducting science experiments on the surface of Mars, communicating with ground control on Earth over time delay.

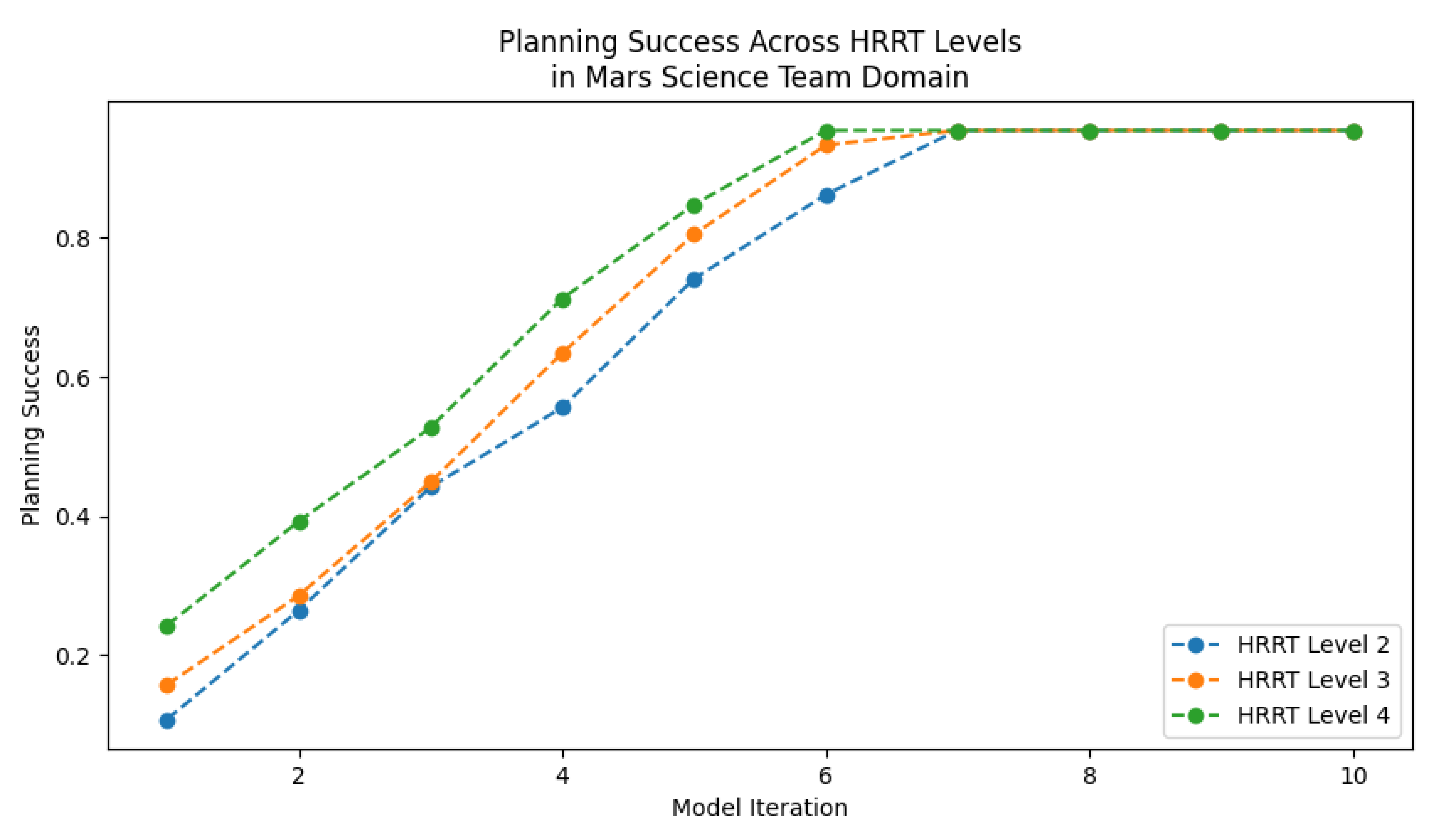
HRRT4: Analysis of "Unknown Unknowns"

Red Computational Agent Prompt	Human-Robot Blue Team Response
Are there additional tasks the robot should be taking on?	Yes, microbial contamination detection and solar panel cleaning.
What catastrophic failures could occur in this domain?	Major data loss or structural damage due to environment.

Computational agents algorithmically generate or look up information in response to prompts, and the human agents determine relevance.

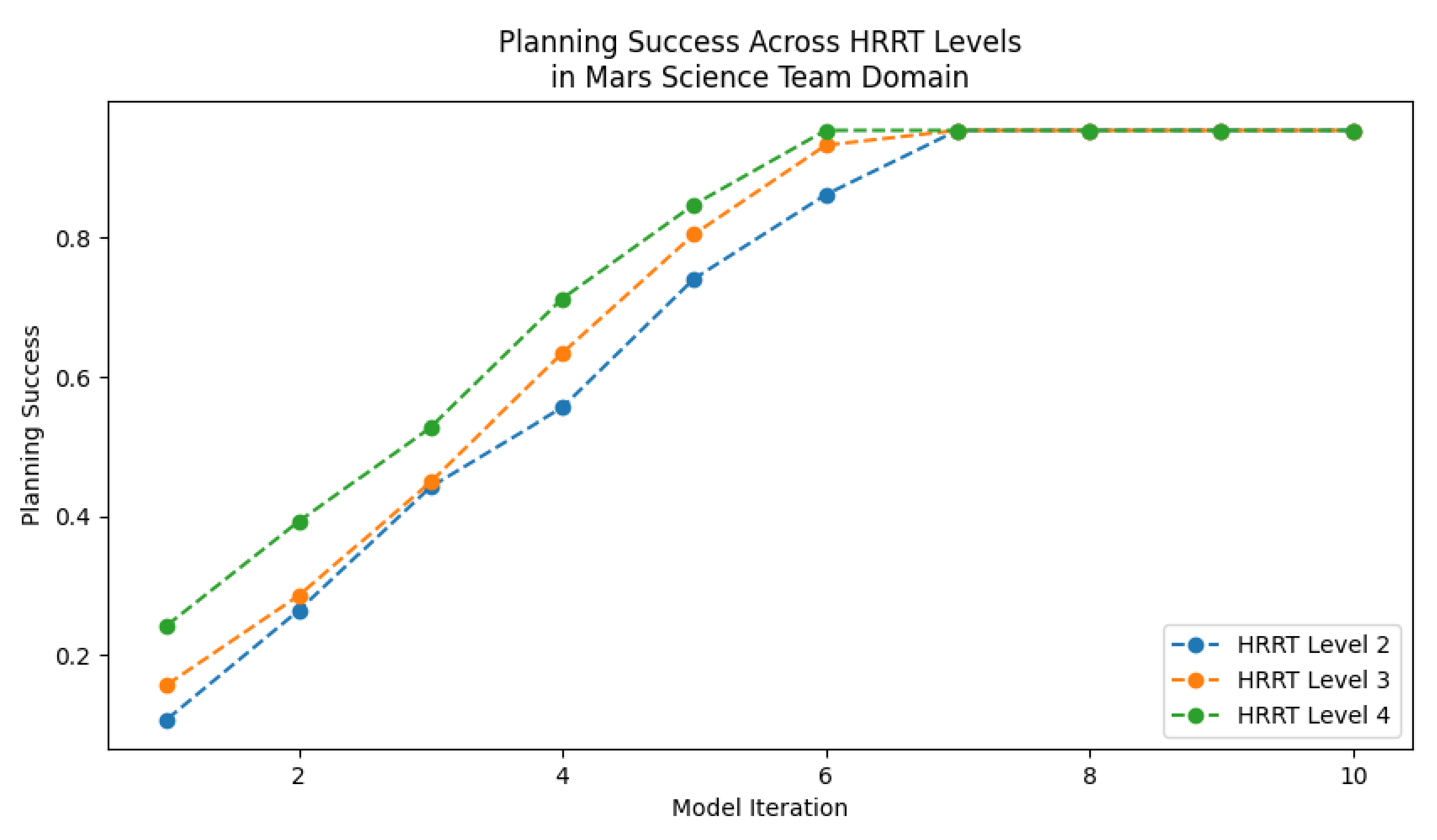
[48] E. Sheetz, "Tool-Use Robot Manipulation Tasks for Cooperative and Explainable Operations in Safety-Critical Domains," PhD Dissertation, University of Michigan, 2025.

Ablation Study over HRRT Levels



Each ablation excludes higher levels of analysis. We tested each model hypothesis in 200 randomized planning tasks, where each task included a random set of failure cases.

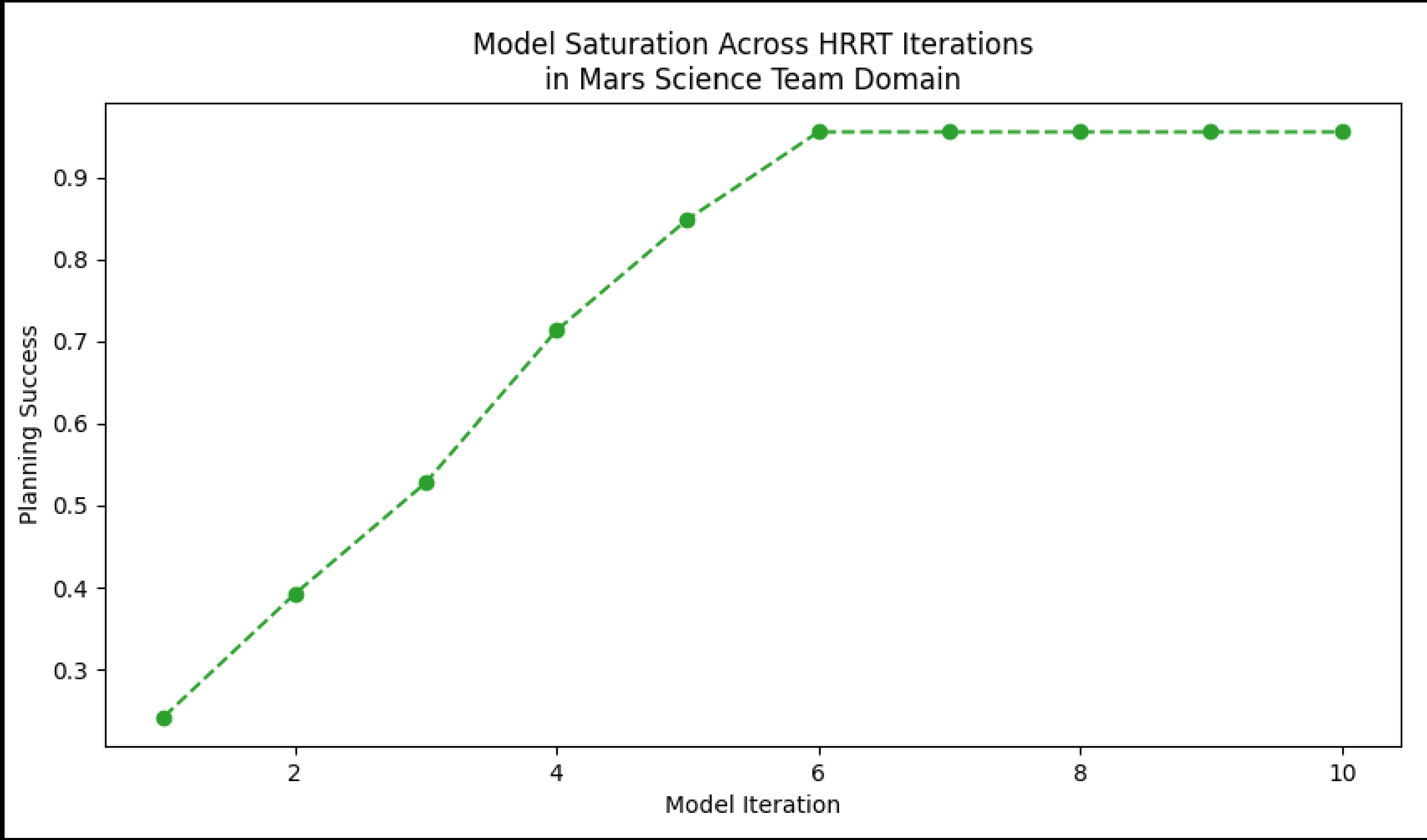
Ablation Study over HRRT Levels



Each HRRT level builds upon the knowledge gained from previous levels. This evidence justifies our iterative process through the interrelated HRRT level analyses.

Model Saturation through HRRT Iterations

These experiments also demonstrate saturation of modeled knowledge through HRRT iterations. After iteration 6, the model contained sufficient risk mitigation mechanisms to plan safely according to our set of failures.

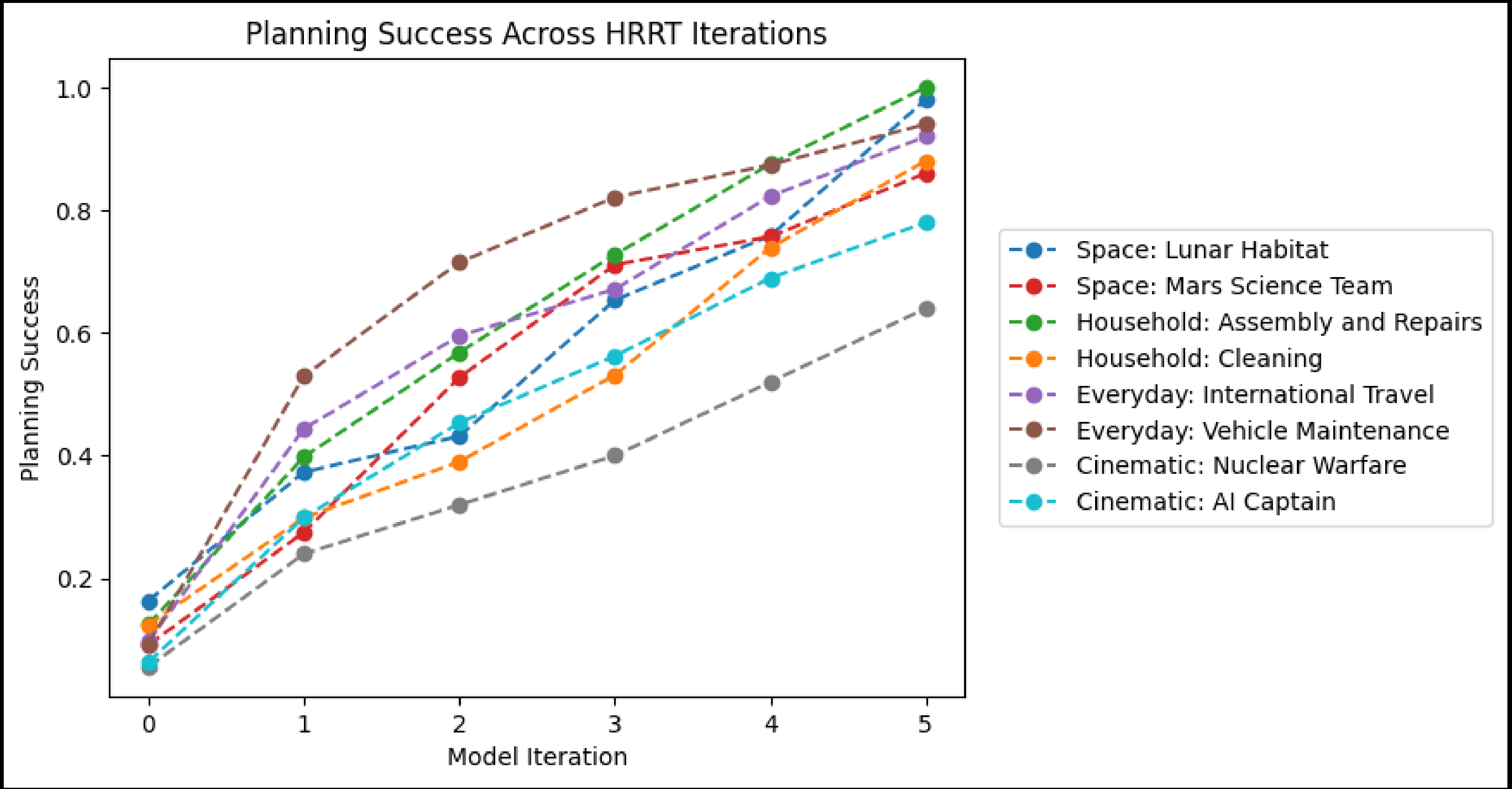


Safety-Critical Planning Domains

- Space Applications
  - Lunar Habitat: assist astronauts in pressurized lunar habitat
  - Mars Science Team: science experiments by team of robots
- Household Applications
  - Assembly and Repairs: regular home maintenance
  - Cleaning: clean a house where family, children, and pets live
- Everyday Applications
  - International Travel: robot personal assistant plans a trip
  - Vehicle Maintenance: robot helps diagnose vehicle issues
- Cinematic Applications
  - Nuclear Warfare: inspired by *The Iron Giant*
  - AI Captain: inspired by *2001: A Space Odyssey*

[49] *The Iron Giant*, Directed by Brad Bird, Warner Bros., 1999.
[50] *2001: A Space Odyssey*, Directed by Stanley Kubrick, Stanley Kubrick Productions, 1968.

Safety-Critical Planning Experiments



Across all tested domains, each iteration made the generated model hypotheses more capable of achieving task goals, mitigating risks, and avoiding critical failures.

Safety-Critical Planning Experiments

Application Class	Problem Domain	Planning Successes	Total Tasks	Success Rate
Space	Lunar Habitat	49	50	0.98
	Mars Science Team	43	50	0.86
Household	Assembly/Repairs	50	50	1.00
	Cleaning	44	50	0.88
Everyday	International Travel	46	50	0.92
	Vehicle Maintenance	47	50	0.94
Cinematic	Nuclear Warfare	32	50	0.64
	AI Captain	39	50	0.78
TOTAL		**350**	**400**	**0.875**

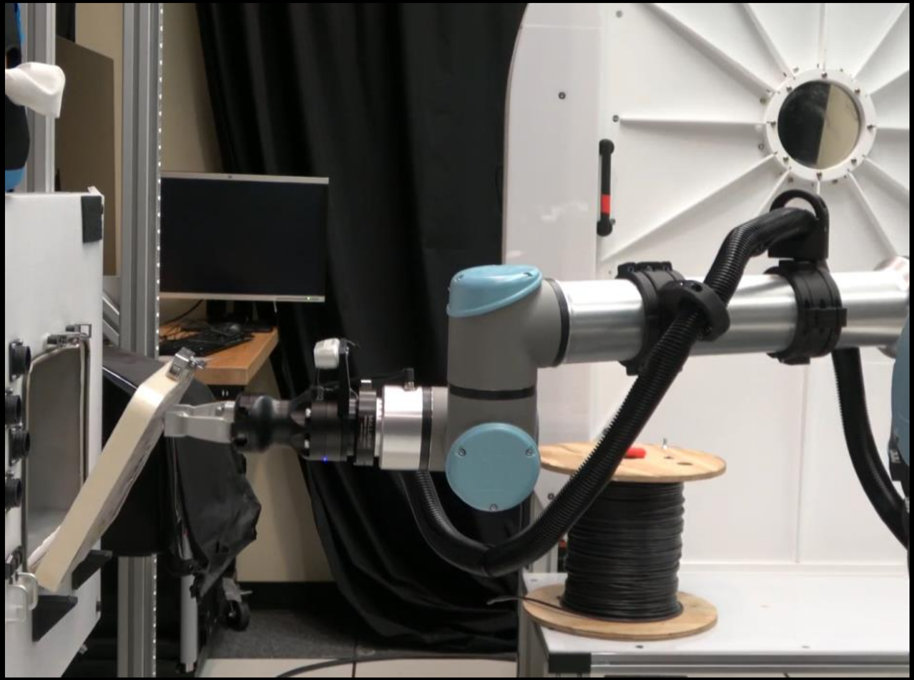
Overall, our HRRT methods help human-robot teams explore complexities of mitigating risks and acting safely.

[51] E. Sheetz *et al.*, "Human-Robot Red Teaming for Safety-Aware Reasoning," *Ubiquitous Robotics*, Under Review, 2025.

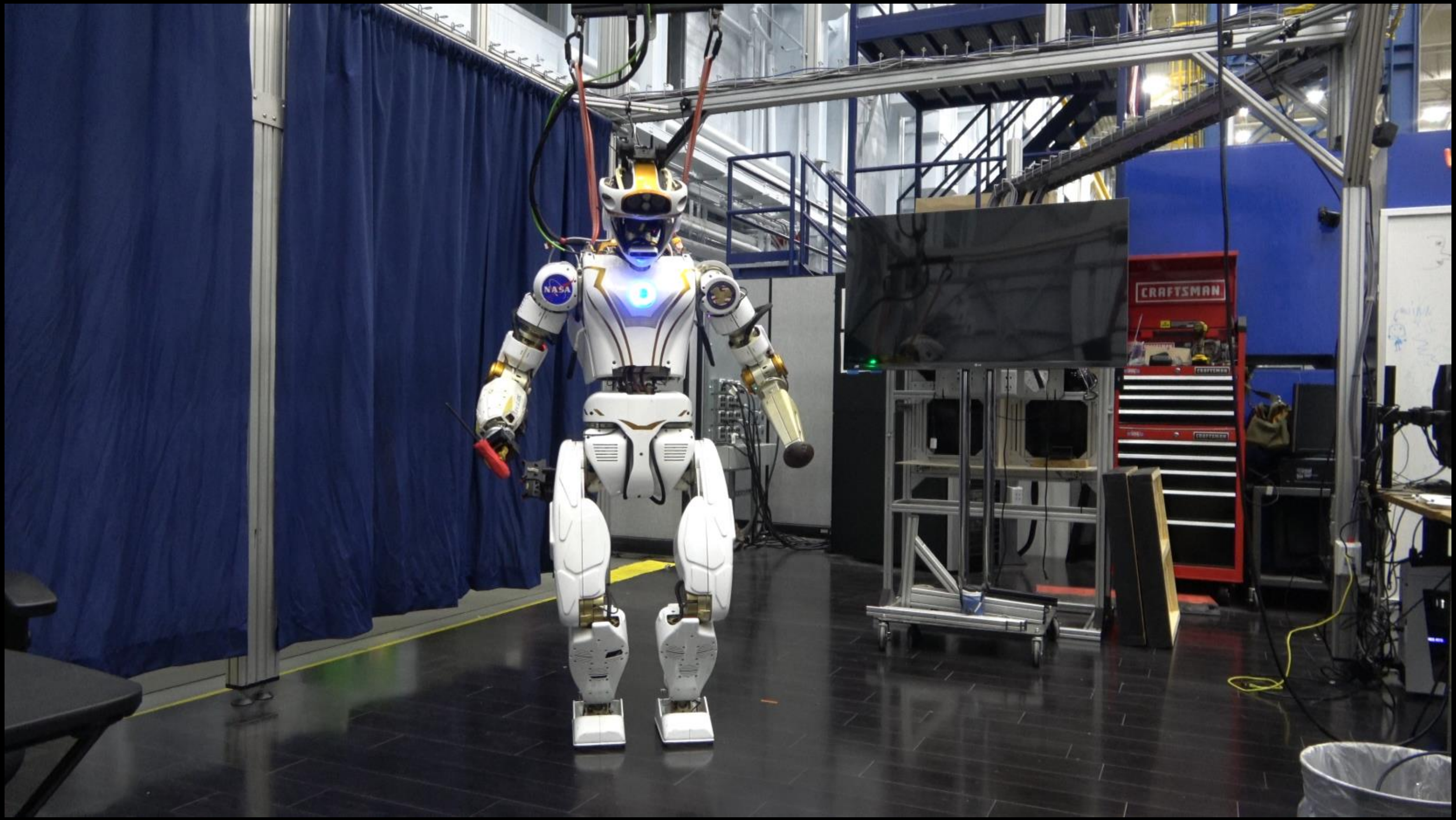
Safety-Critical Execution Experiments

The robots learn to predict the best risk mitigating action based on the data generated by the human-robot red team.

We trained statistically significant, environment-specific risk assessment models for a lunar habitat and household environment.

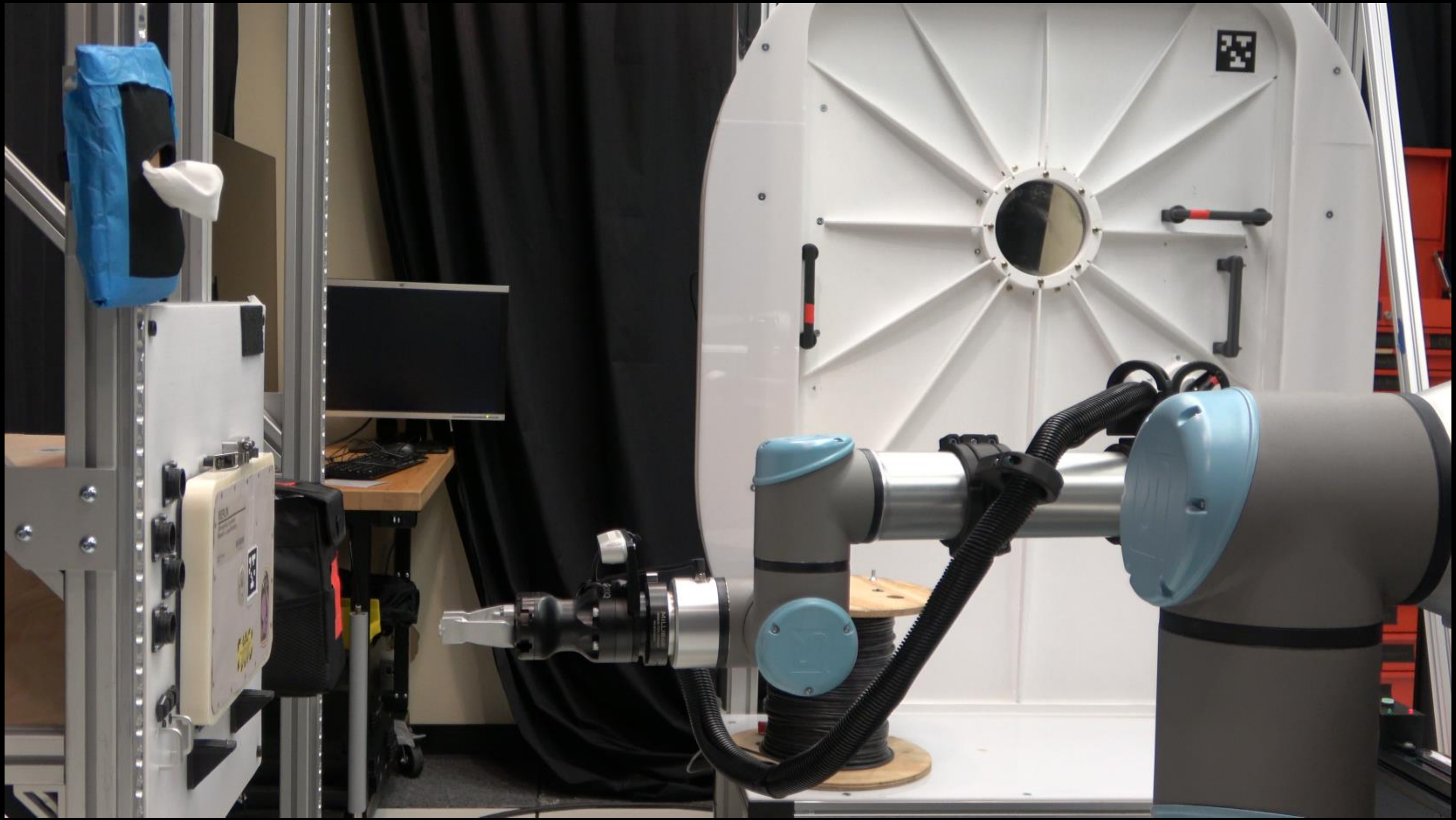


Household Risk Mitigation



The Valkyrie humanoid robot performs tool hand-off tasks in a household environment. Valkyrie assesses the risk of a human walking through the workspace and mitigates the risk by slowing motion to lower risk of injury.

Lunar Habitat Risk Mitigation



The iMETRO armed robot performs sample stowage tasks as if in a lunar habitat. iMETRO assesses the risk of not detecting the sample where expected and mitigates the risk by asking for assistance to complete the task.

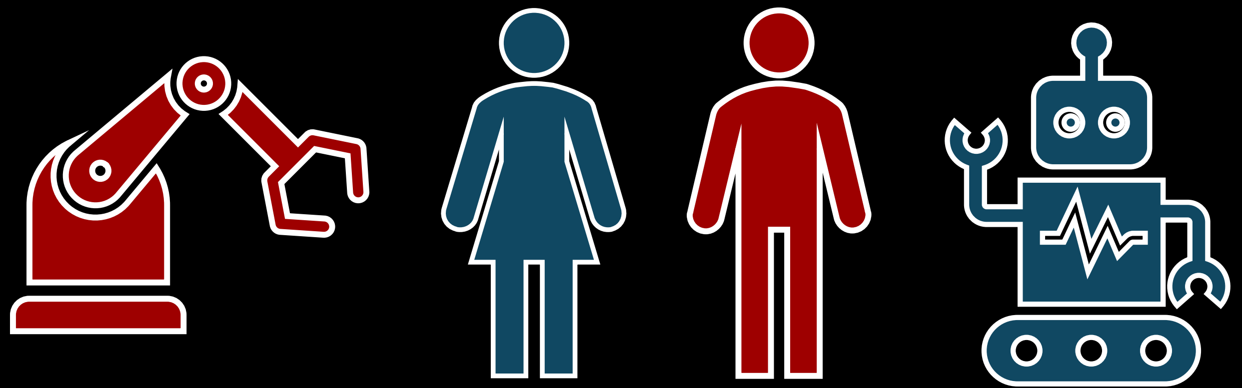
Risk Assessment and Mitigation Results

Environment	Robot	Total Trials	Correct Risk Mitigating Action Success Rate
Lunar Habitat	iMETRO	7	1.00
Household	Valkyrie	5	0.60
Cumulative	**-**	**12**	**0.83**

Robots of different embodiments learned to assess and mitigate risks under different environment-specific definitions of safety through human-robot red teaming.

[51] E. Sheetz *et al.*, "Human-Robot Red Teaming for Safety-Aware Reasoning," *Ubiquitous Robotics*, Under Review, 2025.

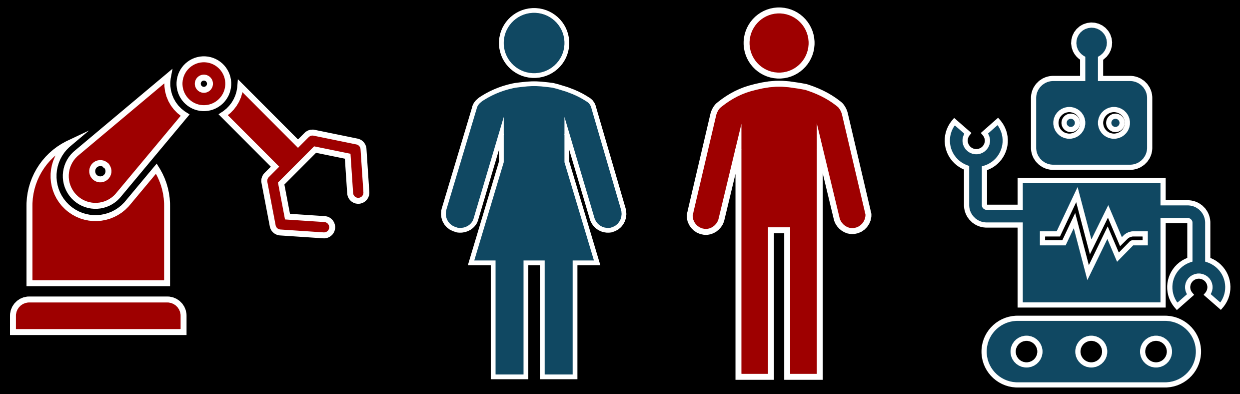
Human-Robot Teams in Safety-Critical Tasks



A complete model M^* of an unboundedly complex world is intractable. A simplified model M makes reasoning possible but may dangerously oversimplify.

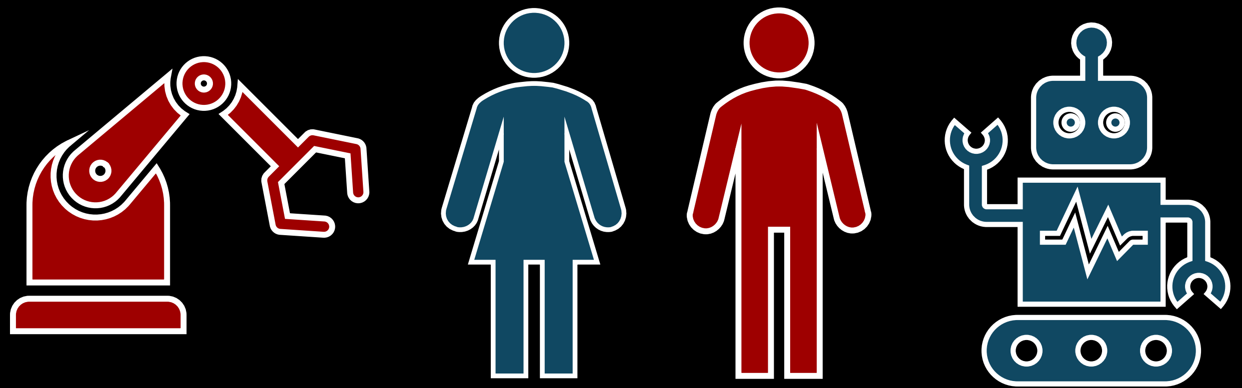
Computational agents have their model M built in, limiting their understanding to symbols in that model. But with their large amount of real-world experience, humans can introduce new symbols to expand the team's understanding to "unknown unknowns."

Human-Robot Teams in Safety-Critical Tasks



Our human-robot red teaming paradigm leverages this diversity of perspectives: robots use computational approaches to systematically challenge the human agents, and humans use their experience to introduce ideas and make evaluative moral judgments.

Human-Robot Teams in Safety-Critical Tasks



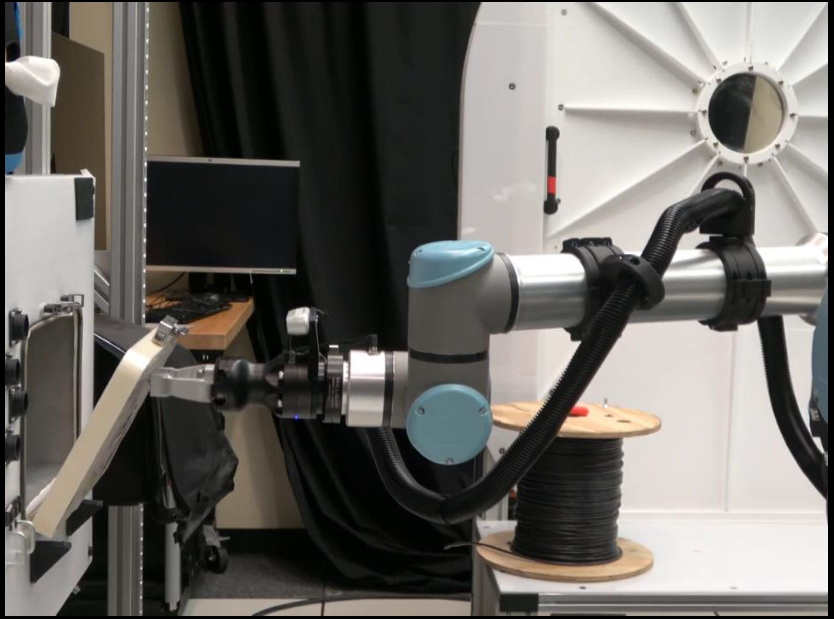
Through this collaborative dialogue, the team iterates on models M, M', M'', ... to improve their ability to plan around and mitigate risks, while still simplifying reasoning over intractable complete model M^*.

The problem of "unknown unknowns" can never be completely solved. But human-robot red teaming provides more opportunities for the team to reason about safety, promote understanding, calibrate trust, and improve knowledge of the problem domain.

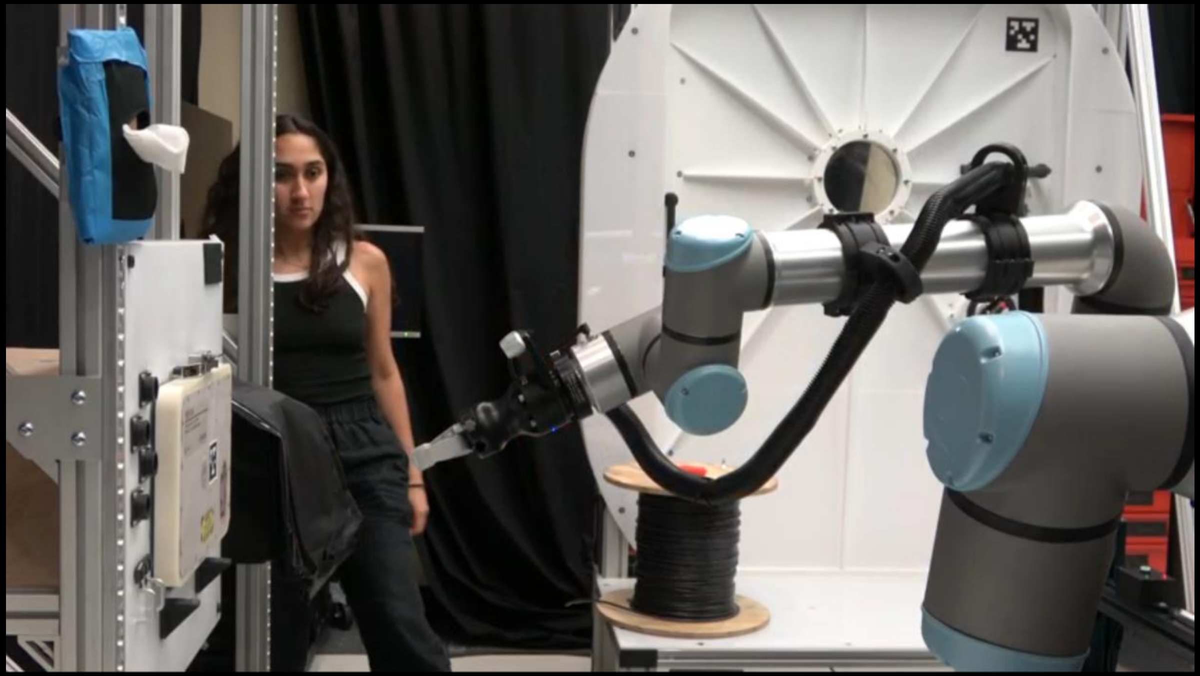
Future Work for Human-Robot Red Teaming

Future work beyond the scope of the dissertation includes:

- Deploying model hypotheses on robots executing real-world tasks
- Investigating composition of human-robot teams for expert insights
- Testing more advanced language capabilities for improved safety dialogue



Safety Reasoning on Human-Robot Teams



The human-robot red teaming approach demonstrates the value of safety reasoning where teams engage in multiple levels of critical analysis in a problem domain (UR 2025, Under Review).

Our methods reduce overtrust and allow robots to earn appropriately calibrated trust on cooperative human-robot teams.

[51] E. Sheetz *et al.*, “Human-Robot Red Teaming for Safety-Aware Reasoning,” *Ubiquitous Robotics*, Under Review, 2025.

Dissertation Contributions

Autonomous planning of complex assembly actions

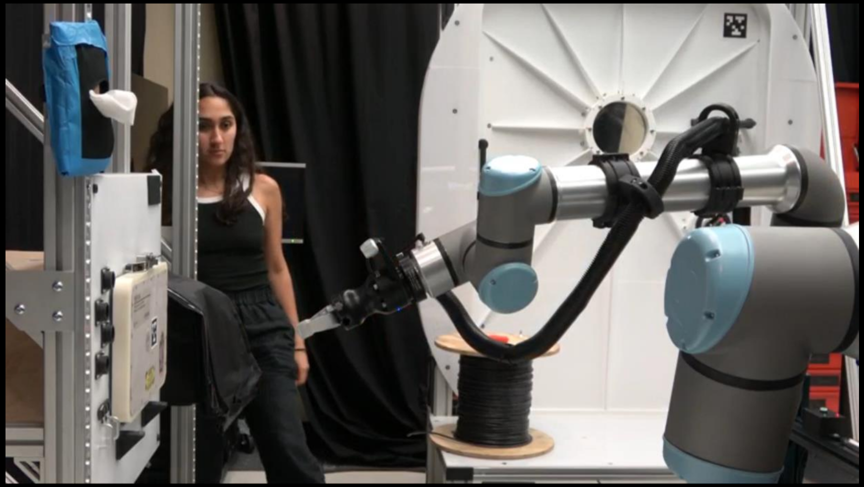
(ICRA 2022)

Reliable and explainable execution of tool-use tasks

(IROS 2024)

Safety reasoning on human-robot teams

(UR 2025, Under Review)

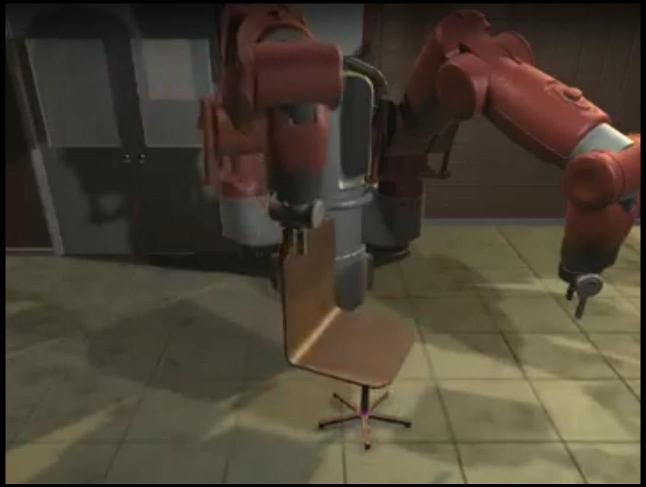


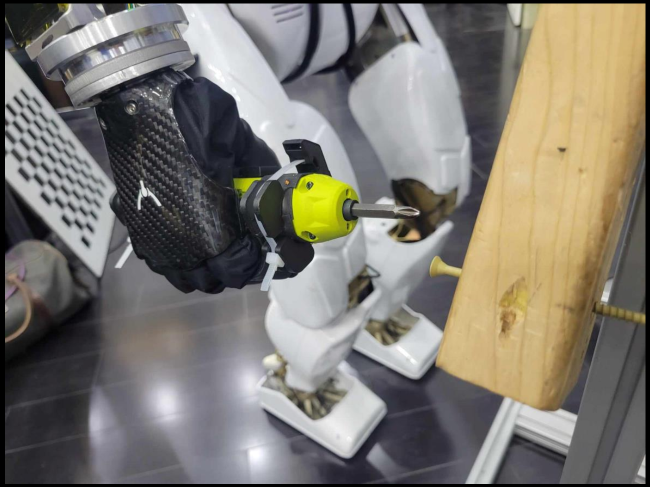
Dissertation Contributions

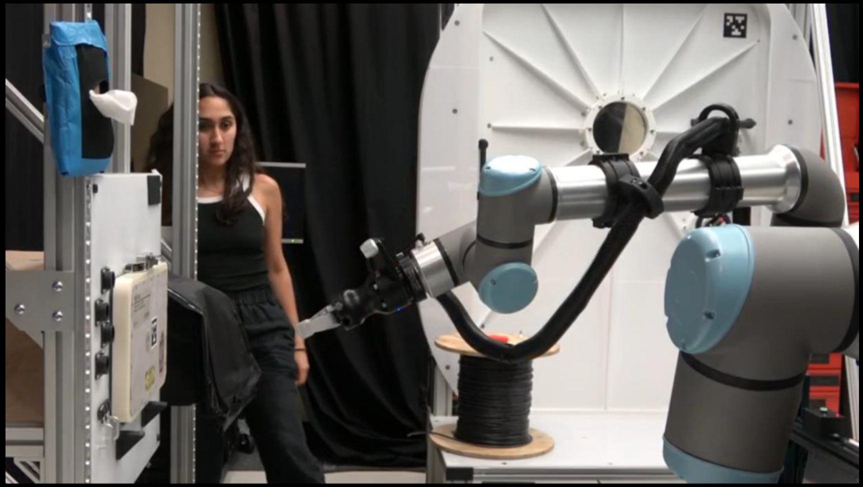
Autonomous planning of complex assembly actions
(ICRA 2022)

Reliable and explainable execution of tool-use tasks
(IROS 2024)

Safety reasoning on human-robot teams
(UR 2025, Under Review)



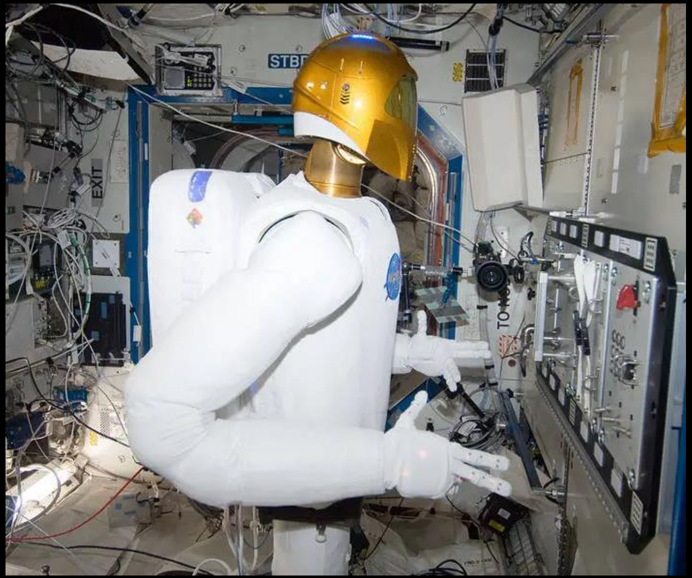


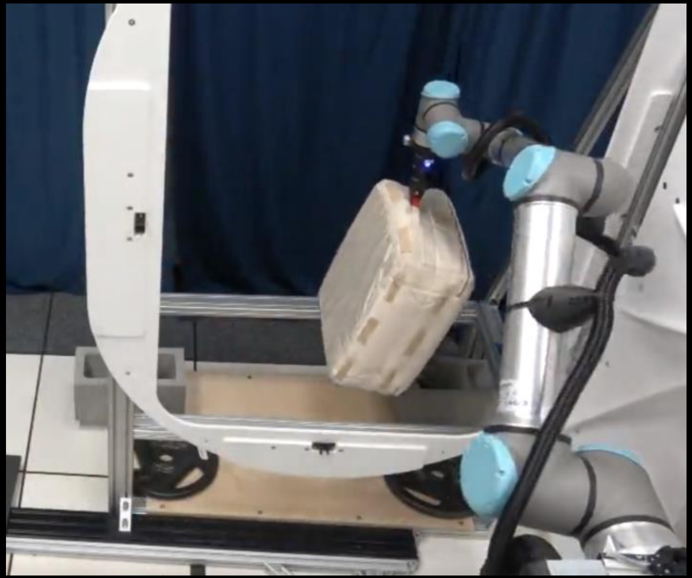


Human-Robot Teams in Safety-Critical Domains

The dissertation explores challenges in (1) robot manipulation capabilities and (2) robot safety reasoning.

Our work contributes to robots operating as capable, trusted agents on human-robot teams in safety-critical problem domains.





Acknowledgements

Committee, Department, Funding

- Benjamin Kuipers
- Chad Jenkins
- Joyce Chai
- Kimberly Hambuchen
- CSE and Robotics Departments, College of Engineering
- NASA Space Technology Graduate Research Opportunity (NSTGRO)
- NASA Johnson Space Center Pathways Internship
- TRACLabs Internship
- CAT Vehicle REU and REU on Smart UAVs

Paper Co-Authors

- **Causal Control Basis**: Xiaotong Chen, Zhen Zeng, Kaizhi Zheng, Qiuyu Shi, and Chad Jenkins
- **Grasp Reflex Model**: Misha Savchenko, Emma Zemler, Abbas Presswala, Andrew Crouch, Shaun Azimi, and Benjamin Kuipers
- **Human-Robot Red Teaming**: Emma Zemler, Misha Savchenko, Connor Rainen, Erik Holum, Jodi Graf, Andrew Albright, Shaun Azimi, and Benjamin Kuipers

Thanks to Mina Kian for her appearance in experiment photos and videos

NASA Johnson Space Center

Division Management:
Stephen Frederickson and Kimberly Hambuchen

Branch Management: Jonathan Rogers

NSTGRO Research Collaborators: Joshua Mehling and Shaun Azimi

Mentors: Misha Savchenko, Emma Zemler, Mark Paterson

Dexterous Robotics Team

Fellow ER4 Interns

Mentors

Professor Michael Sostarecz

Professor James Logan Mayfield

Professor Jonathan Sprinkle

Professor Saad Biaz

Dr. Audra Baleisis

Dr. Stephen Hart

Dr. Ana Huamán Quispe

Dr. Steven Jorgensen

Colleagues and Friends

Teresa French
Emma Vanderpool
Ethan Hager
Ryne Krejci
Emily Danes
Andrew Danes
Heather Ide
Veronica Sells

Sanjay Singapuram
Armand Behroozi
Hafiz Sheriff
Tamara Nelson-Fromm
Dennis Nikolov
Liz Olson
Jorge Vilchis

Tyler Sypherd
Nate Hamilton
Katie Harrold
Andrea Ventola
Mina Kian
Jim Kramer
Amy Fritz
Lauren Nilsson

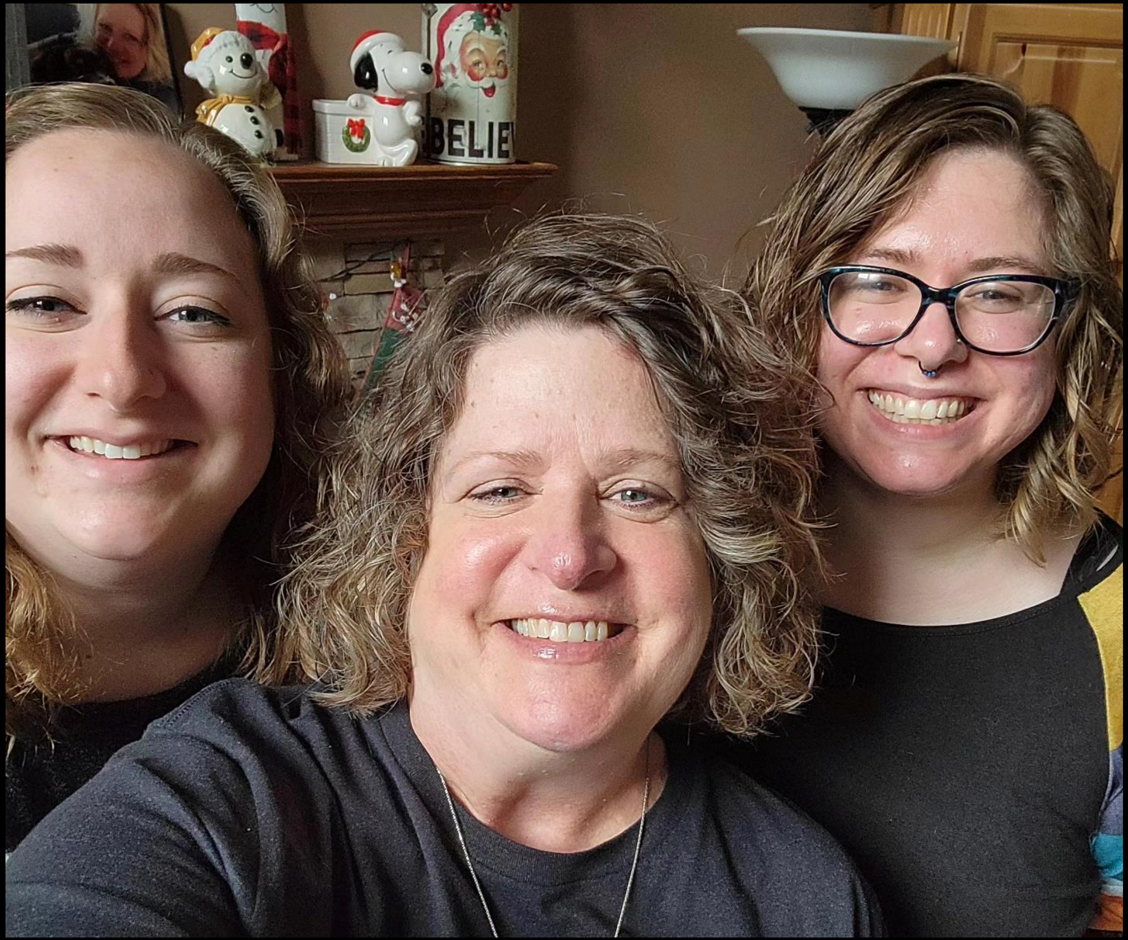
Family

Rita and Tony Tomczak

Mickey Sheetz

Vicky Sheetz

Family



Family

Princess

Boogie

Thank You!

Questions?