



# Explainable AI for Establishing Shared Expectations During Human-Robot Collaboration

Collaborative Artificial Intelligence and Robotics Lab



Prof. Brad Hayes

Bradley.Hayes@Colorado.edu

<http://www.cairo-lab.com/>

<http://www.circadence.com/>

 @hayesbh

 <http://bradhayes.info>



# Focus: Human-Machine Teaming

Chief Technology Officer  
Circadence Corporation

Assistant Professor of Computer Science  
University of Colorado Boulder

Director  
Collaborative AI and Robotics Lab

---

Ph.D. Computer Science, **Yale**

Postdoc, **MIT**

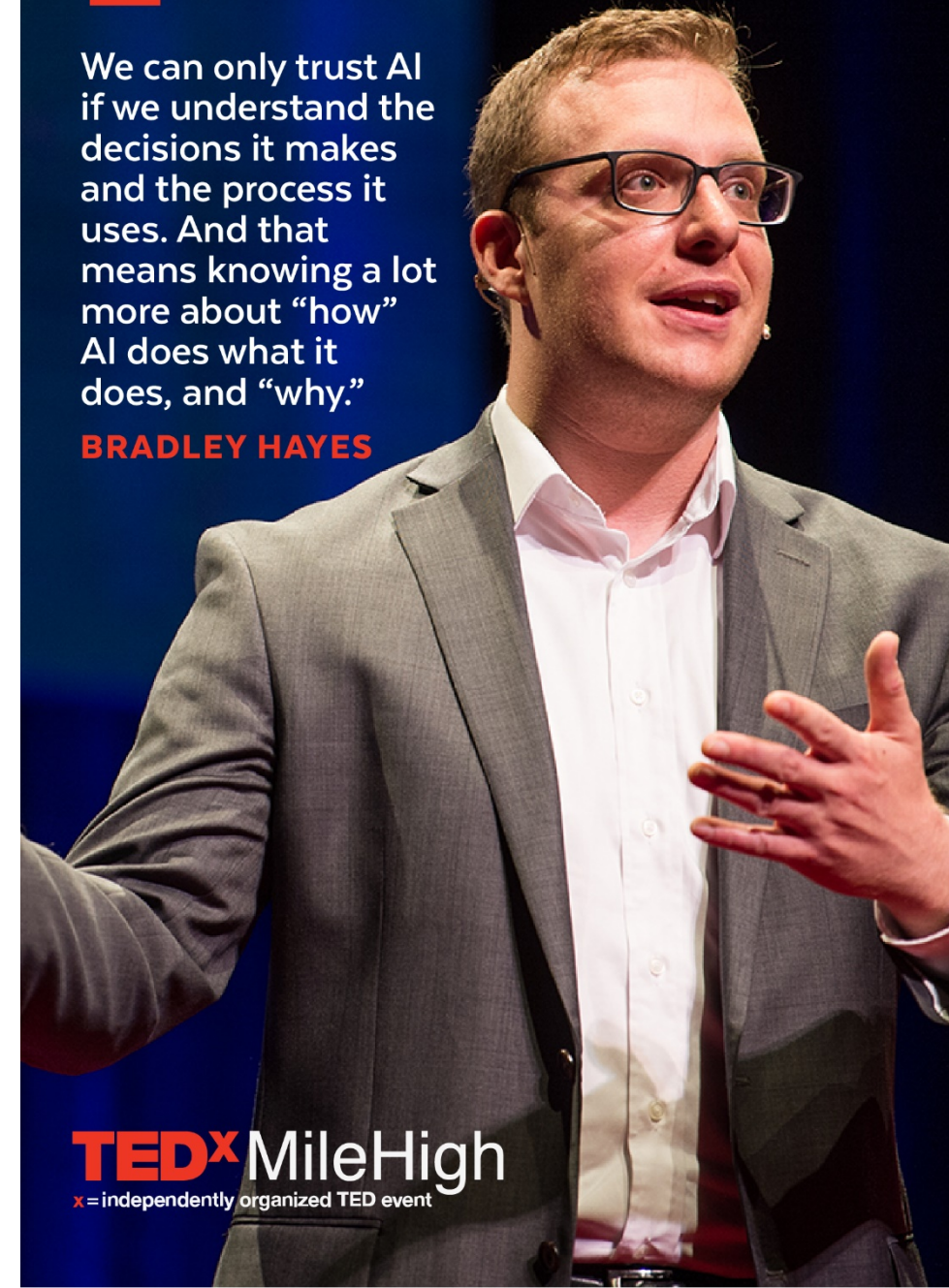


BAE SYSTEMS



We can only trust AI if we understand the decisions it makes and the process it uses. And that means knowing a lot more about “how” AI does what it does, and “why.”

**BRADLEY HAYES**



**TEDx** MileHigh  
x = independently organized TED event





CIRCADENCE

Circadence delivers market leading, immersive, virtual environments for cyber awareness and learning.

Additional focus on operational tools to help cyber defenders defeat evolving threats.



# PROJECT ARES™

NEXT GENERATION CYBERSECURITY TRAINING

**WORLD MAP**  
**CYBER MISSIONS**

ZOOM OUT

SELECT A MISSION TO BEGIN

**MISSION 04**  
Stop Malicious Processes

**MISSION 03**  
Intercept Attack Plan

**MISSION 10**  
**DEFEND HOSPITALS AGAINST RANSOMWARE**  
Prevent malware infections and retain hospital systems for the next two hours to ensure patients who are critical can receive the care they need.

NETWORK SIZE	MEDIUM
COMPLEXITY	SPEC
DEFENSIVE / ONE SIDED	

**MISSION 02**  
Stop Terrorist Financing

**REGION**  
WHOLE GLOBE

- NORTH AMERICA 3
- SOUTH AMERICA 2
- EUROPE 3
- 2 Stop Terrorist Financing
- 4 Stop Malicious Processes
- 10 Defend hospitals against Ransomware
- ASIA 4

**MISSIONS TYPE FILTERS**

- TACTICS
- NETWORK
- COMPLEXITY

**EXPERT**  
XP 10000  
LEVEL 10

PROFILE  
SCORE

**1 NEW MESSAGE**

**YOU**  
Is everyone here? What's up?

**PLAYER2**  
Hey, what's up?

**ATHENA**  
Hi everyone, I am Athena, how can I help you to get started?

**YOU**  
Just completed Mission 4, it was awesome!

**PLAYER2**  
Nice, Congratulatory!

**ATHENA**  
Thank you for completing that!

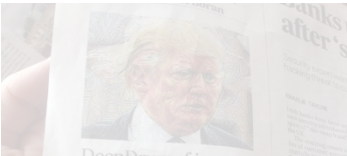
Log out / Hide chat

**CHAT** **PLAYERS**

**PROJECT ARES®**  
NEXT GENERATION CYBERSECURITY TRAINING

INVITATIONS CONFIG HELP

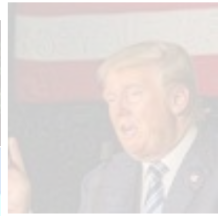




CSAIL at MIT @MIT\_CSAIL · 1m  
 In 5 days, the @DeepDrumpf Trumpbot has surpassed us in followers w/almost 17K. [bit.ly/1TcwtAy](http://bit.ly/1TcwtAy) #Karma



**Deep Drumpf: the Twitter bot trying to out-Trump the Donald**  
 The Guardian - 4 Mar 2016  
 MIT project uses ...  
 MIT built a Donald Trump AI Twitter bot that sounds scarily like ...  
 Quartz - 4 Mar 2016  
**DeepDrumpf** Twitter Bot Pretty Good at Generating Trumpisms  
 AdAge.com (blog) - 4 Mar 2016



Newsweek

**How to Troll Trump and Fundraise for Good, Simultaneously**  
 Inverse - 18 Oct 2016  
 In March, MIT robotics researcher Bradley Hayes conjured up a Donald Trump-emulating Twitter bot: @DeepDrumpf. Now, Hayes is throwing ...

How the AI Behind Twitter's Odd @DeepDrumpf Is Making ...  
 Inverse - 4 Mar 2016

How An AI Donald Trump Is Making Twitter Great Again

Highly Cited - Popular Science  
 Drumpf Twitterbot learns: Blog - CNET (blog) - 31 Mar 2016

[View all](#)



**Halperin12**  
 2:16 PM EST

**This is what scientists do? I want my money back.**

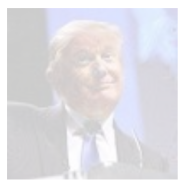
Like Reply Share

designed to tweet like US bot uses ...

What Could ...

DeepDrumpf which spouted tweets

such as "I'll bring back our jobs. They all have everything, ..."



Why MIT Just using



**'DeepDrumpf' Is An Uncanny Twitterbot That's Fundraising For ...**  
 Forbes - 19 Oct 2016  
 DeepDrumpf can be fairly described as a Twitterbot, but it's become a lot more than that over its several months on the 2016 campaign trail.



**An MIT Scientist Created A Trump Twitter Bot And It's Scarily ...**  
 Paste Magazine - 24 Mar 2016

In the beginning, @DeepDrumpf's tweets were pure nonsense; now, they at least resemble coherent statements—though obviously, given the ...



never quite know what  
 Of course, that's what i  
 3/7/2016  
 ZDNet  
 JUST IN MICRO

**Twitterbot uses AI algorithm to tweet like Donald Trump**  
 Matching the real thing for arrogance may be beyond science's grasp.

By Greg Nichols for Robotics | March 4, 2016 -- 11:03 GMT (03:03 PST) | Topic: Robotics

Het Twitteraccount @deepdrumpf is gemaakt door wetenschapper Bradley Hayes van universiteit MIT. Het profiel wordt bestuurd door kunstmatige intelligentie, dat Trumps toespraken heeft geanalyseerd om patronen te herkennen.

Op basis van die analyse doet de 'twitterbot' zelf uitspraken die Trump volgen hebben kunnen doen, zoals bijvoorbeeld "I'm what ISIS doesn't need" (Ik ben red.).



Donald Trump at a rally in Florence, South Carolina, on Feb. 5, 2016  
 Credit: Gage Skidmore/Trump Campaign

Brandon Hayes, Brian Blaney and 2 others

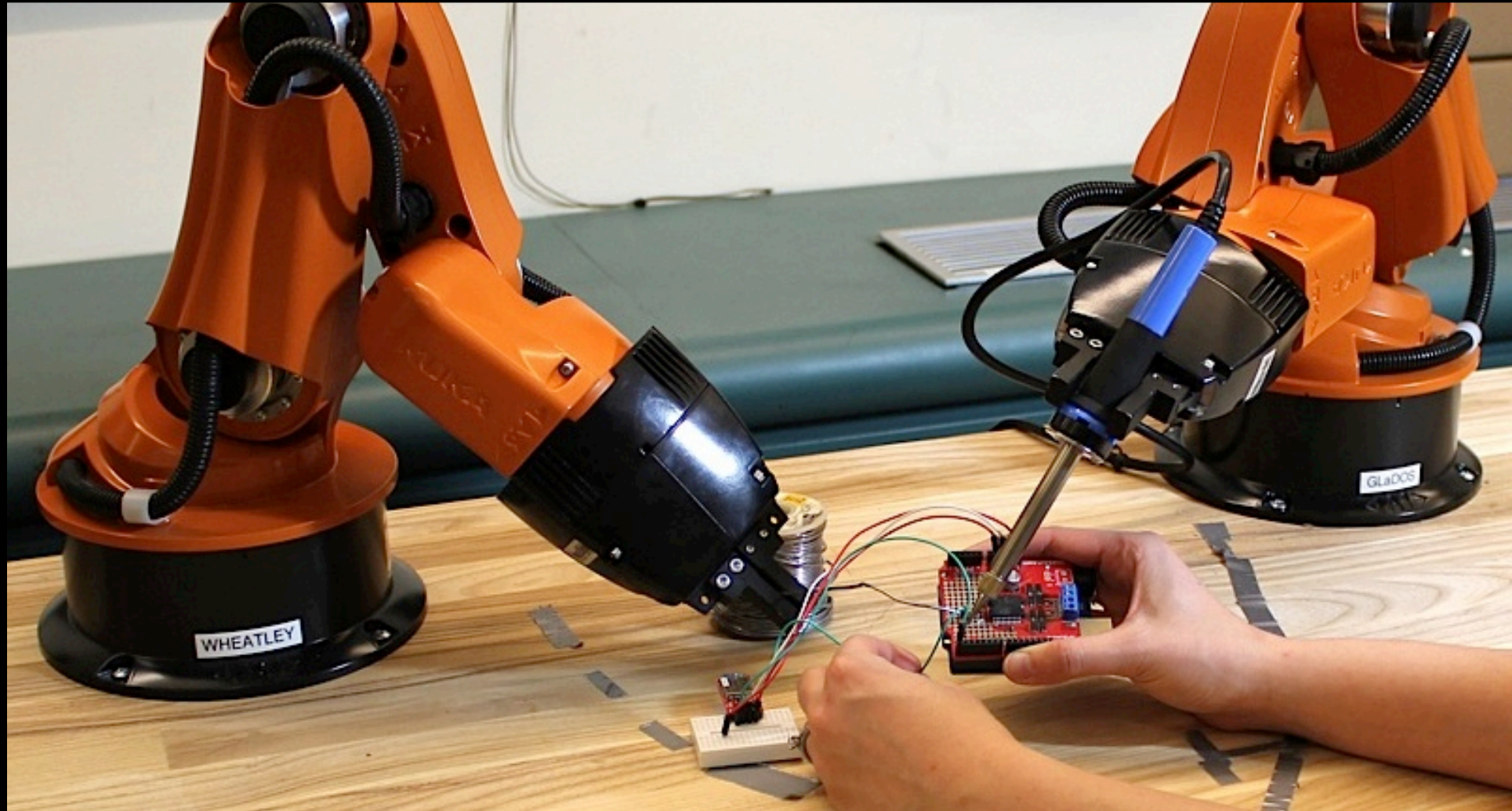
**John Gallagher** I may have never been more proud of a former student  
 Like Reply · 1 hr

5 COM

Write a comment...



# Collaborative Human-Robot Interaction



Human-in-the-loop artificial intelligence enables robot workers to make human collaborators **safer**, more **effective**, and more **efficient**.

# Collaborative Robotics

“Happy People Smiling With Robots”



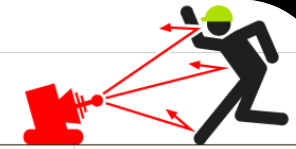
Cages are being replaced by **algorithms**, **sensors**, and **HRI**



# Robot Co-workers



## Robot Accidents in the Workplace



**WHEN:**  
AUGUST 2011

**WHERE:**  
BAKERY

**WHAT HAPPENED:**  
An employee was repairing a jammed conveyor belt in an oven when he became caught between a robotic arm and the belt. He was killed.

**WHEN:**  
MAY 2007

**WHERE:**  
PLASTICS FACTORY

**WHAT HAPPENED:**  
An employee was troubleshooting a robotic arm used to remove CD jewel cases when the arm struck the employee in his head and ribs. He died two weeks later.

**WHEN:**  
JULY 2006

**WHERE:**  
METAL FACTORY

**WHAT HAPPENED:**  
An employee was crushed between a robotic arm and the robot's work station. He appeared to have been reaching to remove a scrap the robot had dropped or to push the reset button, but there was no memory in the robot computer to know for sure. The employee was killed.

**WHEN:**  
MARCH 2006

**WHERE:**  
CAR FACTORY

**WHAT HAPPENED:**  
A robot caught an employee on the back of her neck and pinned her head between itself and the part she was welding. She was killed.

**WHEN:**  
DECEMBER 2001

**WHERE:**  
CAR FACTORY

**WHAT HAPPENED:**  
An employee was cleaning at the end of his shift and entered a robot's unlocked cage. The robot grabbed his neck and pinned the employee under a wheel rim. He was asphyxiated.



**WHEN:**  
AUGUST 1999

**WHERE:**  
METAL FACTORY

**WHAT HAPPENED:**  
A maintenance worker climbed a fence to repair a pin in a robot. It was still operating, and he became caught in the machine. He was killed.



**WHEN:**  
NOVEMBER 1996

**WHERE:**  
SPORTING GOODS MANUFACTURER

**WHAT HAPPENED:**  
An employee was using a robot to weld and drill basketball backboards. When he noticed a half-done hole, he manually drilled it. The robot thought that meant the cycle was complete and unexpectedly turned, pinning the employee against the wall. He was hospitalized.

**WHEN:**  
JUNE 1999

**WHERE:**  
MEATPACKING PLANT

**WHAT HAPPENED:**  
An employee accidentally activated a robot when he stepped on a conveyor belt where robots were moving boxes of meat. He became trapped. When his co-workers removed the robot, he fell to the floor. He was killed.



**WHEN:**  
FEBRUARY 1996

**WHERE:**  
ALUMINUM FACTORY

**WHAT HAPPENED:**  
Three workers were watching a robot pour molten aluminum when the pouring unexpectedly stopped. One of them left to flip a switch to start the pouring again. The other two were still standing near the pouring operation, and when the robot restarted, its 150-pound ladle pinned one of them against the wall. He was killed.






**Task Execution**



**Collaborative Task Execution**



An iceberg floating in a blue ocean. The tip of the iceberg is above the water, while the much larger base is submerged. A yellow box with a yellow border is positioned in the upper right, with a yellow line pointing to the tip of the iceberg. The text 'Collaborating During Task Execution' is inside the box. The submerged part of the iceberg has white text overlaid on it.

**Collaborating**  
During Task Execution

Unsolvable?  
Unrelatable?  
Unsafe?



Collaborative

**Collaborating**  
During Task Execution

Understandable

**Shared Expectations:**  
Decision-making

Safe

**Shared Expectations:**  
Behaviors

# Donald Michie's criteria for Machine Learning (ML)

## Weak criterion:

ML occurs whenever a system generates an updated basis building on sample data for improving its performance on subsequent data.

## Strong criterion:

Weak criterion + ability of system to communicate internal updates in explicit symbolic form.

## Ultra-strong criterion:

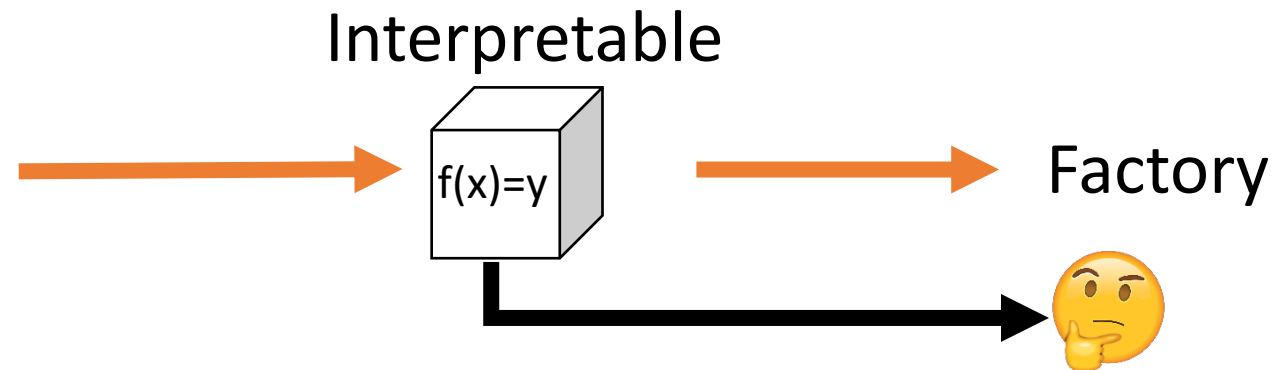
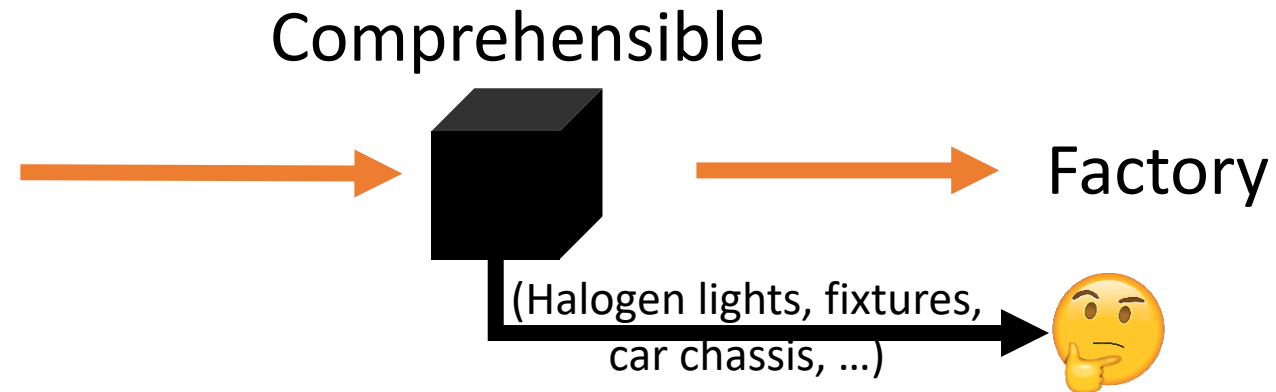
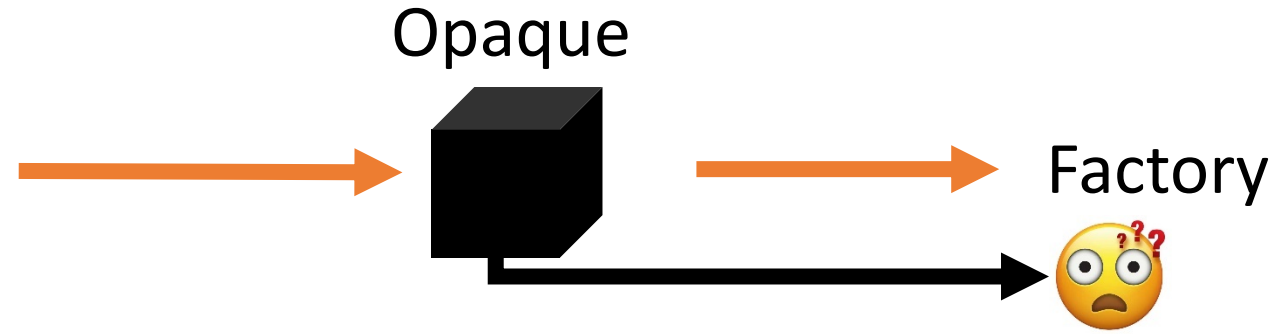
Strong criterion + communication of updates must be operationally effective (i.e. user is required to understand updates and consequences should be drawn from it).







# Relating Different Types of Systems





# Classifying Wolves vs. Huskies





# Learning from Demonstration



...because we aren't very good at crafting cost/reward functions



# Using AI for Road Navigation



# Context-sensitive Assistance Using LfD!





# Learning from Demonstration



...but sometimes we aren't great at demonstrations either

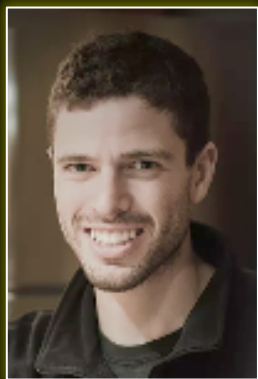


Did the robot *really* capture my intent?



# Robust Robot Learning from Demonstration and Skill Repair Using Conceptual Constraints

[IROS 18]



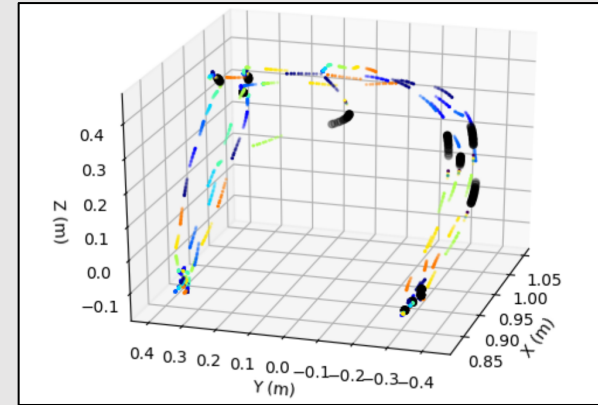
Carl Mueller



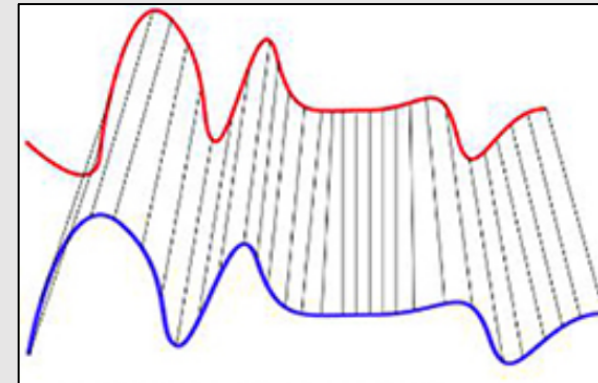
Jeff Venix

# A Typical Learning from Demonstration Pipeline

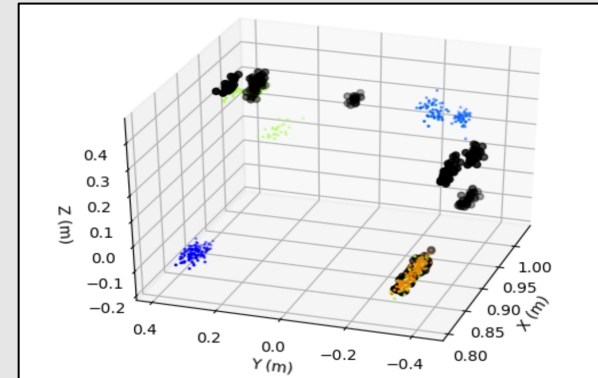
Record Trajectories



Perform Alignment



Cluster and Train Keyframes/Subgoals





# What's Wrong with Learning from Demonstration?

Trajectory-based demonstrations have the lowest overhead  
... but also limited information content.

Implied constraints (e.g., cups should be carried upright) are  
generally drawn from common sense  
... which your robot does not have

'Common Sense' from Demonstration requires a prohibitively  
large number of trajectories

... which you probably don't have time for  
... which you probably have to borrow  
... which you probably still shouldn't trust

# Constrained Task and Motion Planning as a Solution to the “Common Sense Problem”



Orientation e.g., Grasping and holding constraints

Positional e.g., Above, below, around target or obstacle

Motion e.g., Speed or Acceleration



# Constrained Learning from Demonstration

## Key Insights

Narration provides 'common sense' substitute:  
Soliciting and incorporating high level constraints into subgoal execution eases correctness burden from training data

### Increase Skill Robustness

Improves execution under conditions not seen during training

### Reduce Data Requirements

Learns more flexible, generalizable representations with less data

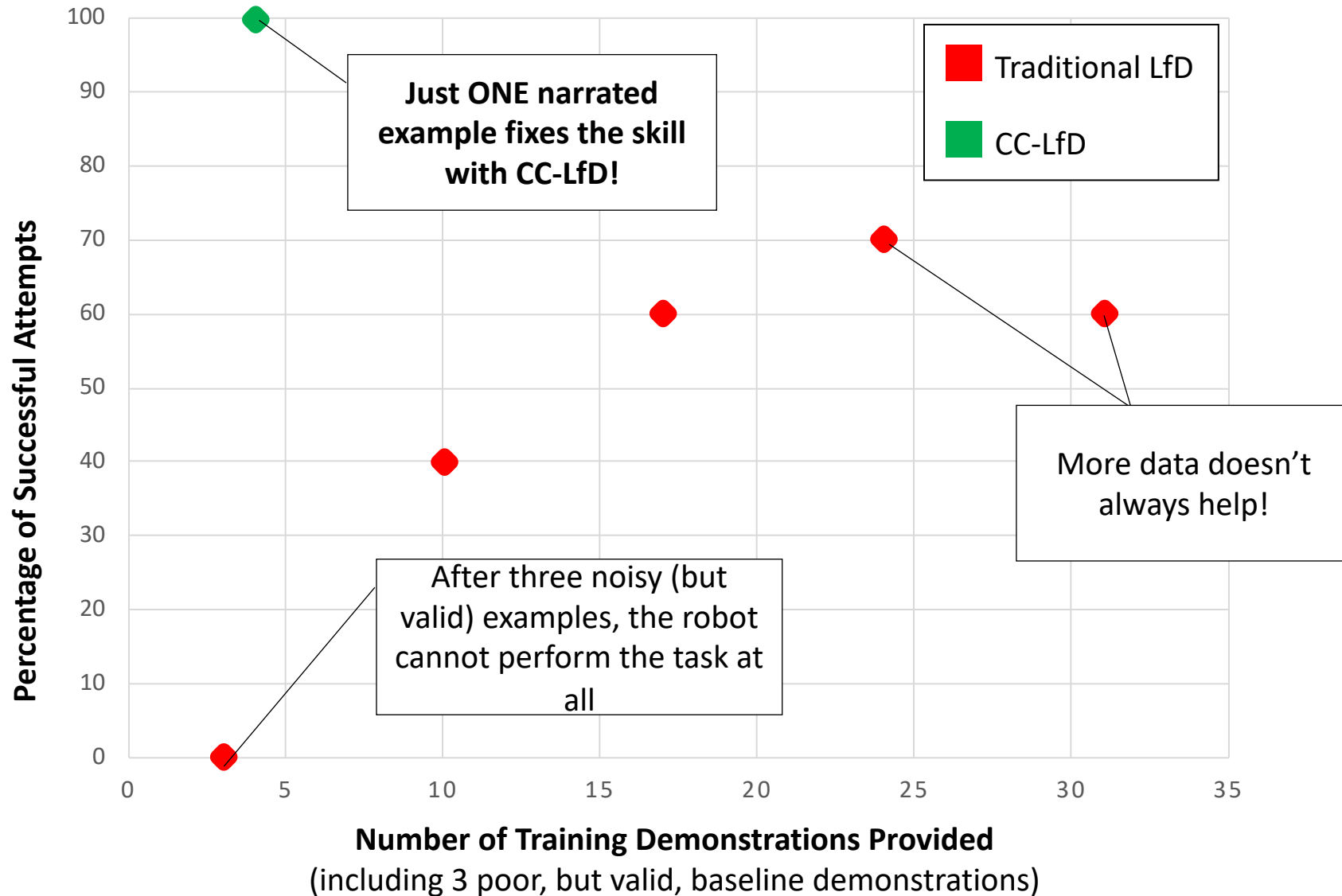
### Increase Resilience to Poor Training

Avoids skill failures even when trained with sub-optimal demonstrations

### Improve and Repair Existing Skills

Enables one-shot skill repair to improve existing skills with a single new example

# CONSTRAINED LEARNING FROM DEMONSTRATION SUCCESS: “POURING TASK” ROBOT PERFORMANCE AND ONE-SHOT SKILL REPAIR





# The Promise of Collaborative Robots



# The Reality of Mismatched Expectations







## Improving Robot Controller Transparency Through Autonomous Policy Explanation

[HRI 17]

# Shared Expectations are Critical for Teamwork

In close human-robot collaboration...

- Humans must be able to plan around robot behaviors
- Understanding failure modes and policies are central to ensuring safe interaction and **managing risk**



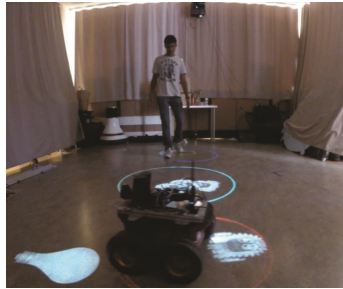
Fluent teaming **requires** communication...

- When there's no prior knowledge
- When expectations are violated
- When there is joint action

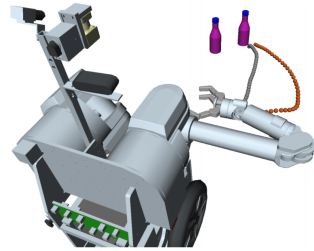




# Establishing Shared Expectations



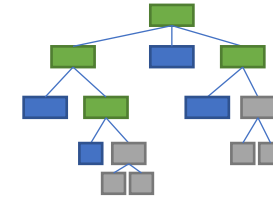
Role-based Feedback  
[St. Clair et al. 2016]



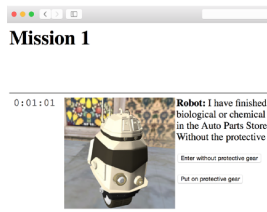
Legible Motion  
[Dragan et al. 2013]

	BREAK	IDLE	NEGOTIATE	SELL	INNERTALK	WATCH	GUARD	EQUIP
ANNY	0	0	0	1	1	1	0	0
BENNY	0	0	0	1	1	1	0	0
CANNY	1	0	0	0	0	0	0	1
DANNY	0	0	1	1	1	0	0	0
ERNY	0	1	0	0	0	0	1	0
FRENNY	1	0	0	0	0	0	0	1

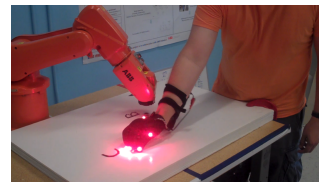
Coordination Graphs  
[Kalech 2010]



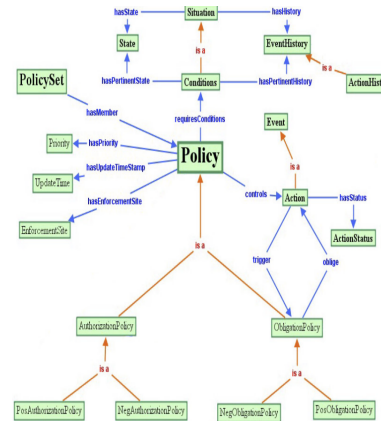
Hierarchical Task Models  
[Hayes et al. 2016]



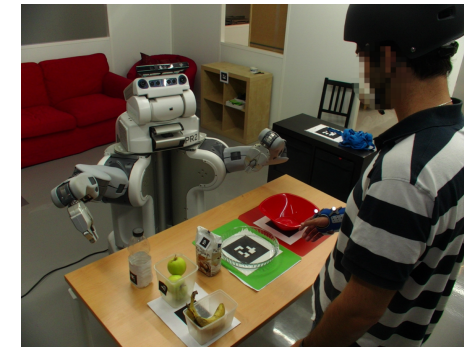
State Disambiguation  
[Wang et al. 2016]



Cross-training  
[Nikolaidis et al. 2013]



Policy Dictation  
[Johnson et al. 2006]



Collaborative Planning  
[Milliez et al. 2016]

Short Term

Long Term

# Semantics for Policy Transfer

Under what conditions  
will you drop the bar?



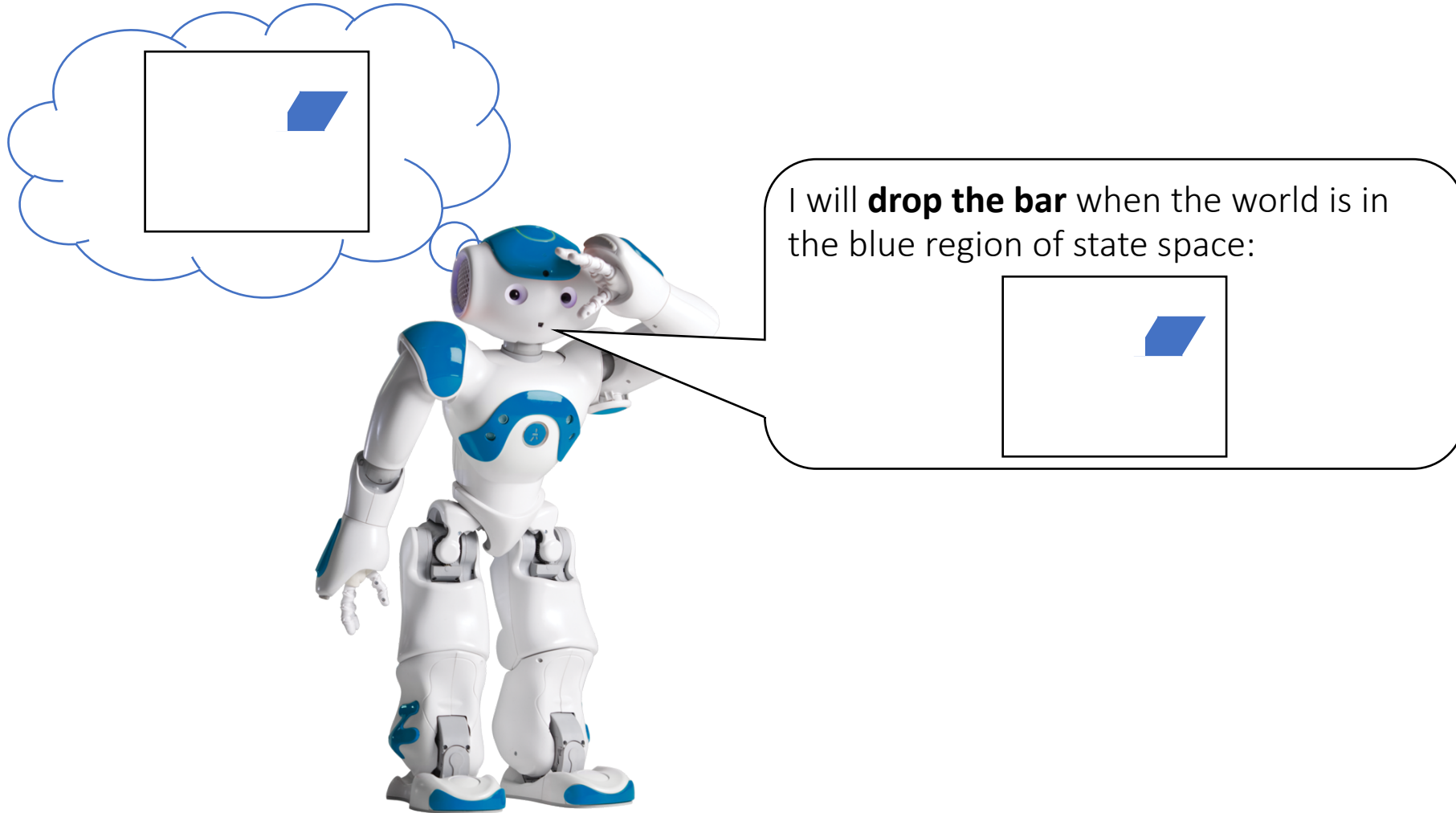


# Semantics for Policy Transfer

Under what conditions  
will you drop the bar?

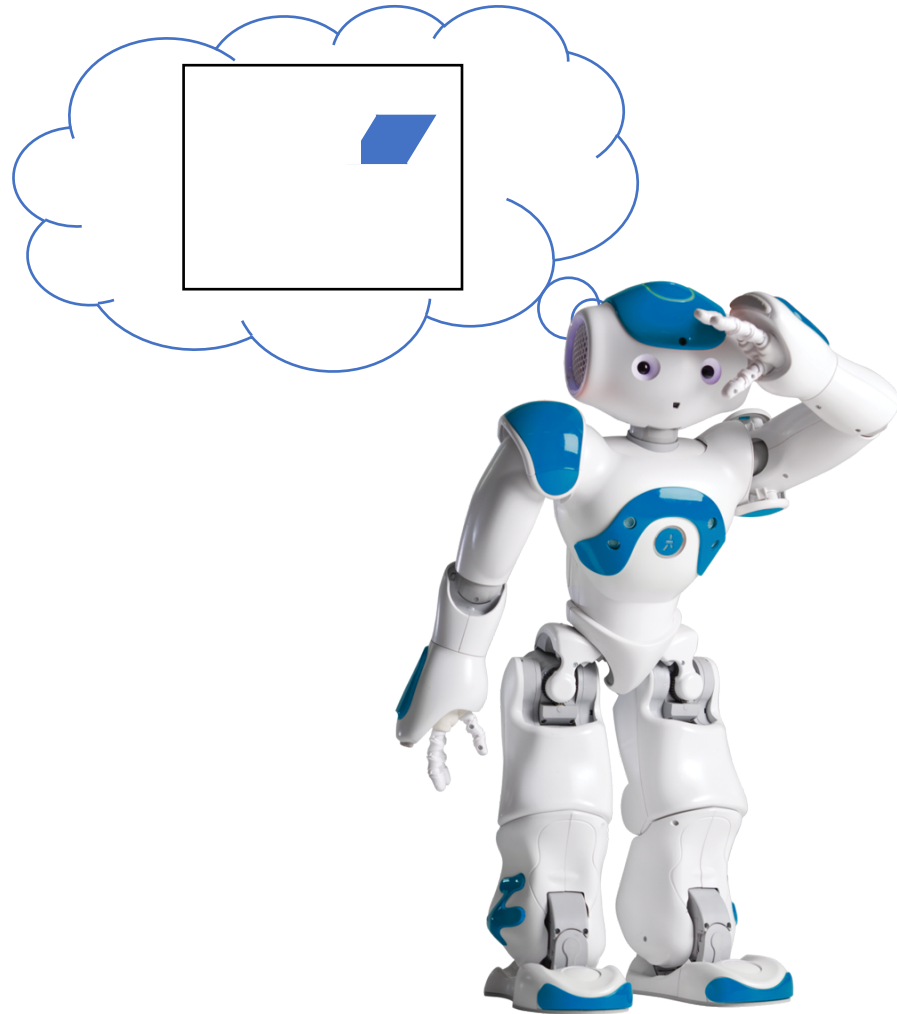


# Semantics for Policy Transfer

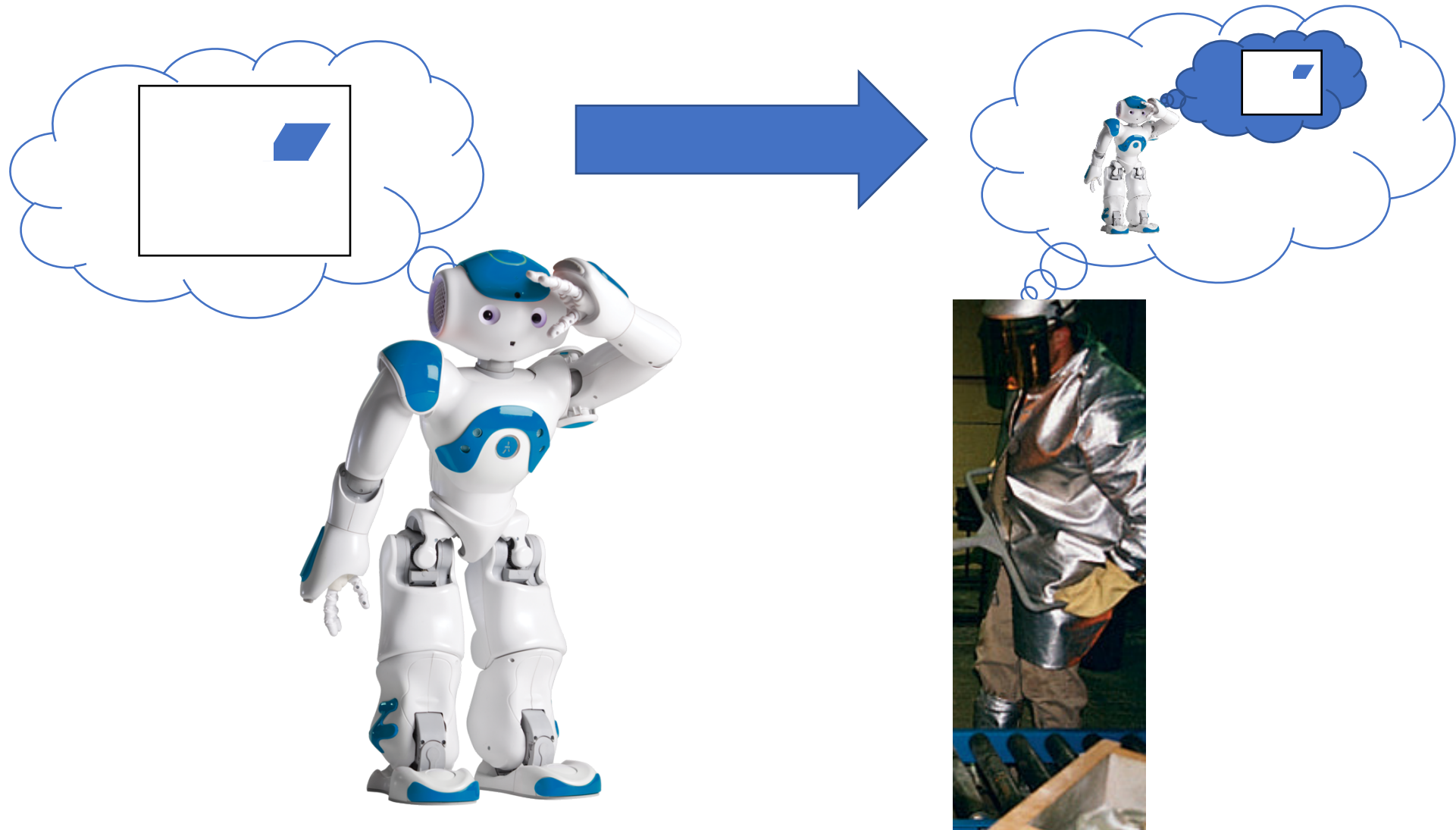




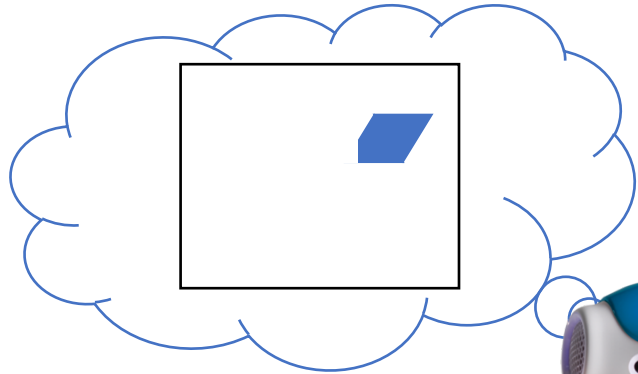
# Semantics for Policy Transfer



# Semantics for Policy Transfer







I will **drop the bar** when the world is in the blue region of state space:

12.4827  
5.12893  
1.12419  
0  
0  
1  
3.62242  
-40.241  
...

15  
7.125  
1.12419  
0  
0  
1  
-8.1219  
-40  
...

12.4827  
8.51422  
1.12419  
0  
1  
0  
3.62242  
-40.241  
...

,

,

...

# State space is too obscure to directly articulate



I will **drop the bar** when the world is in the blue region of state space:

12.4827  
5.12893  
1.12419  
0  
0  
1  
3.62242  
-40.241  
...


15  
7.125  
1.12419  
0  
0  
1  
-8.1219  
-40  
...

12.4827  
8.51422  
1.12419  
0  
1  
0  
3.62242  
-40.241  
...

, , ...

# State of the Art

**Logic Diagram**



**LD**

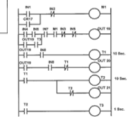
**Description**

- Example Thermal processing of metals. The machine hardens the metal in the shape of a steel ring. The hardening process is done by heating the steel ring to a very high temperature, then it goes through a sudden cooling. So the piece we want to harden is heated by passing very high currents through a coil that heats the piece, then we cool it very quickly by sending cold water through the holes in each side.

**Representation**

- Start/Stop/Switch/  
• Solenoid/Valve/Coil

Symbol	Name	Address	Comment
(S)	Start/Stop	00000	Start/Stop
(S)	Stop	00001	Stop
(S)	Emergency Stop	00002	Emergency Stop
(S)	Left Limit	00003	Left Limit
(S)	Right Limit	00004	Right Limit
(S)	High Stop	00005	High Stop
(S)	High Stop	00006	High Stop
(S)	High Stop	00007	High Stop
(S)	High Stop	00008	High Stop
(S)	High Stop	00009	High Stop
(S)	High Stop	00010	High Stop
(S)	High Stop	00011	High Stop
(S)	High Stop	00012	High Stop
(S)	High Stop	00013	High Stop
(S)	High Stop	00014	High Stop
(S)	High Stop	00015	High Stop
(S)	High Stop	00016	High Stop
(S)	High Stop	00017	High Stop
(S)	High Stop	00018	High Stop
(S)	High Stop	00019	High Stop
(S)	High Stop	00020	High Stop
(S)	High Stop	00021	High Stop
(S)	High Stop	00022	High Stop
(S)	High Stop	00023	High Stop
(S)	High Stop	00024	High Stop
(S)	High Stop	00025	High Stop
(S)	High Stop	00026	High Stop
(S)	High Stop	00027	High Stop
(S)	High Stop	00028	High Stop
(S)	High Stop	00029	High Stop
(S)	High Stop	00030	High Stop
(S)	High Stop	00031	High Stop
(S)	High Stop	00032	High Stop
(S)	High Stop	00033	High Stop
(S)	High Stop	00034	High Stop
(S)	High Stop	00035	High Stop
(S)	High Stop	00036	High Stop
(S)	High Stop	00037	High Stop
(S)	High Stop	00038	High Stop
(S)	High Stop	00039	High Stop
(S)	High Stop	00040	High Stop
(S)	High Stop	00041	High Stop
(S)	High Stop	00042	High Stop
(S)	High Stop	00043	High Stop
(S)	High Stop	00044	High Stop
(S)	High Stop	00045	High Stop
(S)	High Stop	00046	High Stop
(S)	High Stop	00047	High Stop
(S)	High Stop	00048	High Stop
(S)	High Stop	00049	High Stop
(S)	High Stop	00050	High Stop
(S)	High Stop	00051	High Stop
(S)	High Stop	00052	High Stop
(S)	High Stop	00053	High Stop
(S)	High Stop	00054	High Stop
(S)	High Stop	00055	High Stop
(S)	High Stop	00056	High Stop
(S)	High Stop	00057	High Stop
(S)	High Stop	00058	High Stop
(S)	High Stop	00059	High Stop
(S)	High Stop	00060	High Stop
(S)	High Stop	00061	High Stop
(S)	High Stop	00062	High Stop
(S)	High Stop	00063	High Stop
(S)	High Stop	00064	High Stop
(S)	High Stop	00065	High Stop
(S)	High Stop	00066	High Stop
(S)	High Stop	00067	High Stop
(S)	High Stop	00068	High Stop
(S)	High Stop	00069	High Stop
(S)	High Stop	00070	High Stop
(S)	High Stop	00071	High Stop
(S)	High Stop	00072	High Stop
(S)	High Stop	00073	High Stop
(S)	High Stop	00074	High Stop
(S)	High Stop	00075	High Stop
(S)	High Stop	00076	High Stop
(S)	High Stop	00077	High Stop
(S)	High Stop	00078	High Stop
(S)	High Stop	00079	High Stop
(S)	High Stop	00080	High Stop
(S)	High Stop	00081	High Stop
(S)	High Stop	00082	High Stop
(S)	High Stop	00083	High Stop
(S)	High Stop	00084	High Stop
(S)	High Stop	00085	High Stop
(S)	High Stop	00086	High Stop
(S)	High Stop	00087	High Stop
(S)	High Stop	00088	High Stop
(S)	High Stop	00089	High Stop
(S)	High Stop	00090	High Stop
(S)	High Stop	00091	High Stop
(S)	High Stop	00092	High Stop
(S)	High Stop	00093	High Stop
(S)	High Stop	00094	High Stop
(S)	High Stop	00095	High Stop
(S)	High Stop	00096	High Stop
(S)	High Stop	00097	High Stop
(S)	High Stop	00098	High Stop
(S)	High Stop	00099	High Stop



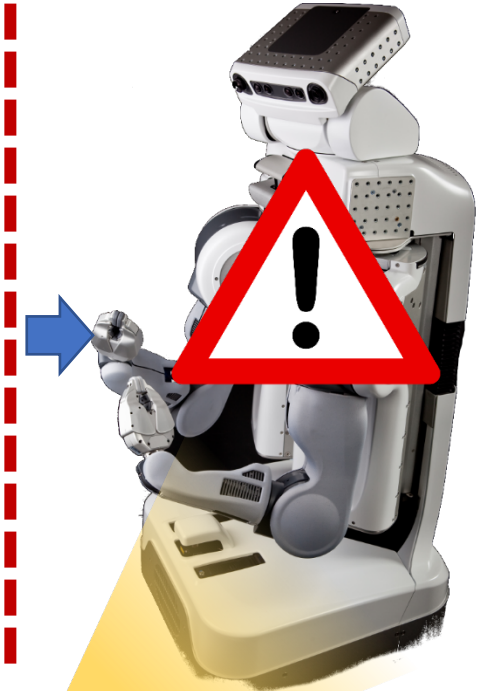
**LD Logic Diagram**

- IN1,2...,M1,CR17...,OUT19...,T1,...



```
int *detect_gear = &INPUT1;
int *gear_x = &INPUT2;

if (*detect_gear == 1 && *gear_x <= 10 && *gear_x >= 8) {
    pick_gear(gear_x);
}
```





# Making Opaque Systems More Understandable in 3 ~~Easy~~ Steps

Computationally Expensive

## Approach:

1. Map human-posed queries to state regions
2. Minimally summarize the identified state regions
3. Communicate query response using natural language

Query Analysis

Response Generation

# Concept Representations

**Concept library:** generic state classifiers mapped to semantic templates that identify whether a state fulfills a given criteria

Set of Boolean classifiers:      State  $\rightarrow$  {True, False}

- Spatial concepts                      (e.g., "A is on top of B")
- Domain-specific concepts          (e.g., "Widget paint is drying")
- Agent-specific concepts            (e.g., "Camera is powered")



on\_top(A,B)



camera\_powered

# General Question Templates

When will you do {action}?



## Algorithm 2: Identify Dominant-action State Region

**Input:** Behavioral Model  $G = \{V, E\}$ , Target Action  $a_t$

**Output:** Set of target states  $S_{\pi^a}$ , Set of non-target states

```
 $S_{\pi^* \setminus a}$ 
1  $S_{\pi^a} \leftarrow \{\}$ ;
2  $S_{\pi^* \setminus a} \leftarrow \{\}$ ;
3 foreach  $s \in V$  do
4    $a \leftarrow$  most frequent action executed from  $s$ ;
5   if  $a == a_t$  then  $S_{\pi^a} \leftarrow S_{\pi^a} \cup s$ ;
6   else  $S_{\pi^* \setminus a} \leftarrow S_{\pi^* \setminus a} \cup s$ ;
7 return  $S_{\pi^a}, S_{\pi^* \setminus a}$ ;
```



# General Question Templates

Why didn't you do {action}?



## Algorithm 3: Identify Behavioral Divergences

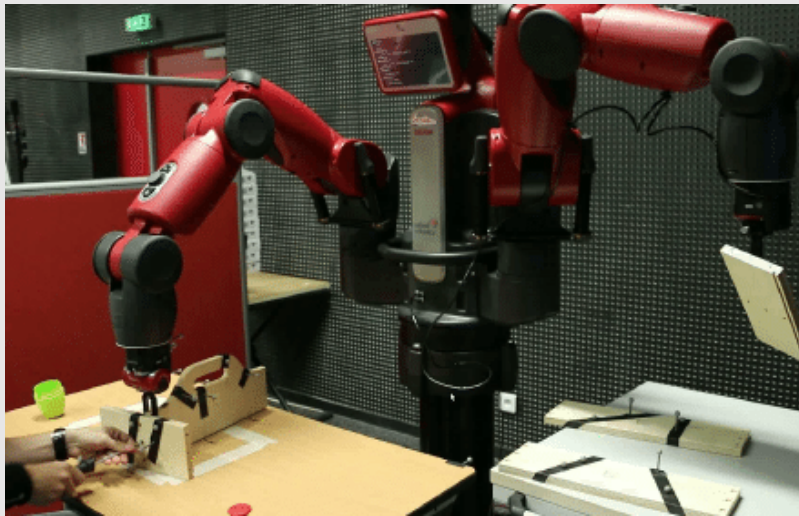
**Input:** Behavioral Model  $G = \{V, E\}$ , Target Action  $a_t$ , Previous state  $s_p$ , Distance threshold  $D_{const}$

**Output:** Explanation of difference between current state and state region where  $a_t$  is performed, explanation of where  $a_t$  is performed locally.

```
1  $S_{\pi^a} \leftarrow \{\}$ ;
2  $S_{\pi^* \setminus a} \leftarrow \{\}$ ;
3 foreach  $D \in \{1, \dots, D_{const}\}$  do
4   foreach  $s \in \{v \in V \mid distance(v, s_p) \leq D\}$  do
5      $a \leftarrow$  most frequent action executed from  $s$ ;
6     if  $a == a_t$  then  $S_{\pi^a} \leftarrow S_{\pi^a} \cup s$ ;
7     else  $S_{\pi^* \setminus a} \leftarrow S_{\pi^* \setminus a} \cup s$ ;
8 expected_region  $\leftarrow$  describe( $G, S_{\pi^a}, S_{\pi^* \setminus a}$ );
9 current_region  $\leftarrow$  describe( $G, \{s_p\}, S_{\pi^a}$ );
10 return diff(expected_region, current_region),
    expected_region;
```

# General Question Templates

What will you do when {conditions}?



## Algorithm 4: Characterize Situational Behavior

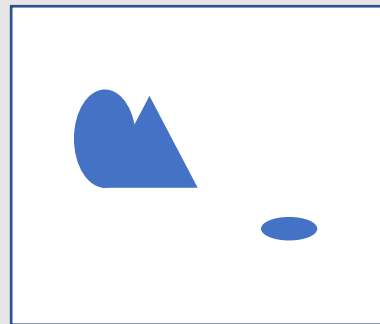
**Input:** Behavioral Model  $G = \{V, E\}$ , Concept Library  $C$ , State region description  $d$ , Max action threshold  $cluster\_max$

**Output:** Explanation of behavior in  $d$ , broken down by action and accompanying state region

```
1  $S \leftarrow dict()$ ;  
2  $descriptions \leftarrow dict()$ ;  
3  $DNF\_description \leftarrow convert\_to\_DNF\_formula(d, C)$ ;  
4 foreach  $s \in \{v \in V \mid test\_dnf(v, DNF\_description) \text{ is } True\}$  do  
5    $S[\pi(s)] \leftarrow S[\pi(s)] \cup s$ ;  
6   if  $|S| > cluster\_max$  then  
7      $\_ return\ too\_many\_actions\_error$   
8 foreach  $a \in S$  do  
9    $descriptions[a] \leftarrow describe(S[a])$ ;  
10 return  $descriptions$ ;
```

# Language Mapping: Model to Response

**Recall:** Concept library provides dictionary of classifiers that cover state regions



on\_top(A,B)

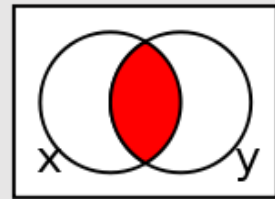


camera\_powered

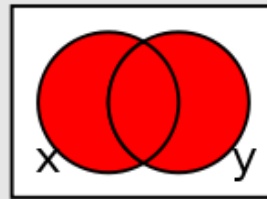


# Using Concepts to Describe State Regions

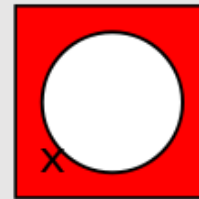
We perform **state-to-language mapping** by applying a Boolean algebra over the space of concepts



$$x \wedge y$$



$$x \vee y$$



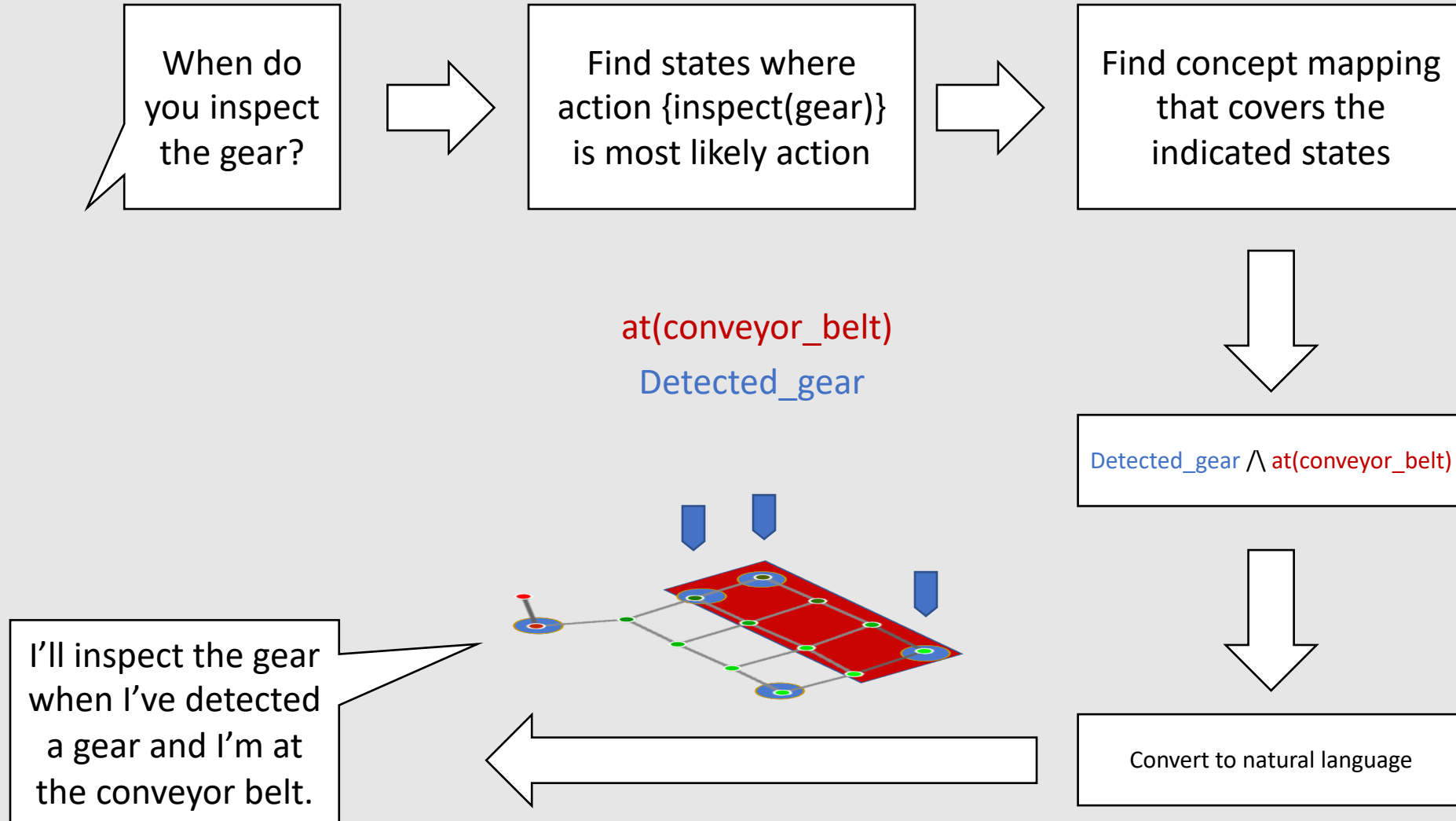
$$\neg x$$

This reduces concept selection to a **set cover problem** over state regions

Disjunctive normal form (DNF) formulae enable coverage over arbitrary geometric state space regions via **intersections** and **unions** of concepts

Templates provide a mapping from DNF  $\rightarrow$  natural language

# Query Response Process



# Explainable Models are not Enough

Interpretability and comprehensibility **enable** explanations,  
but do not yield explanations themselves!

## Reasonable answer:

“My camera didn’t see a gear. I inspect the gear when it is less than 0.3m from the conveyor belt center and it has been placed by the gantry.”

Fault Diagnosis

Policy Explanation

Root Cause Analysis

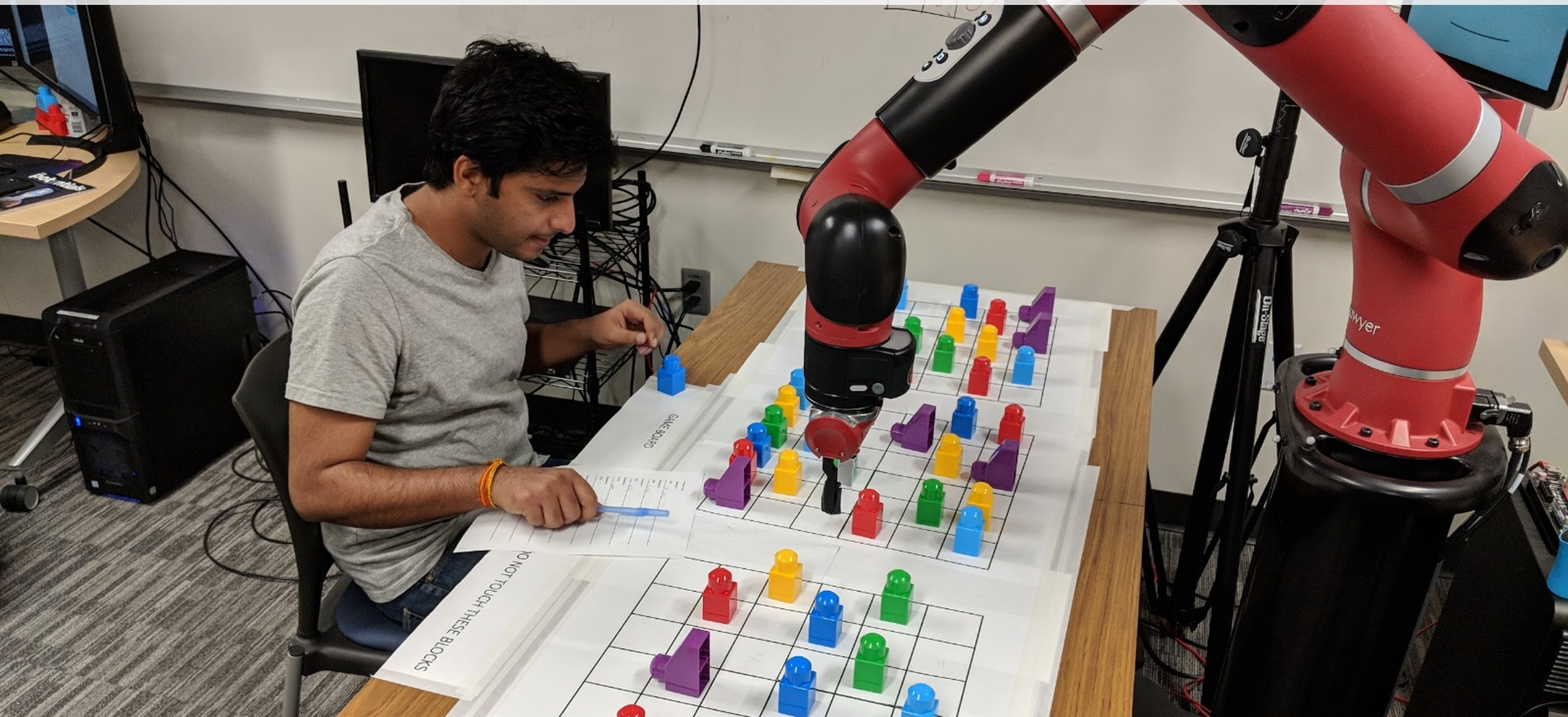


Shaping **Robots** to Match **Human** Expectations!



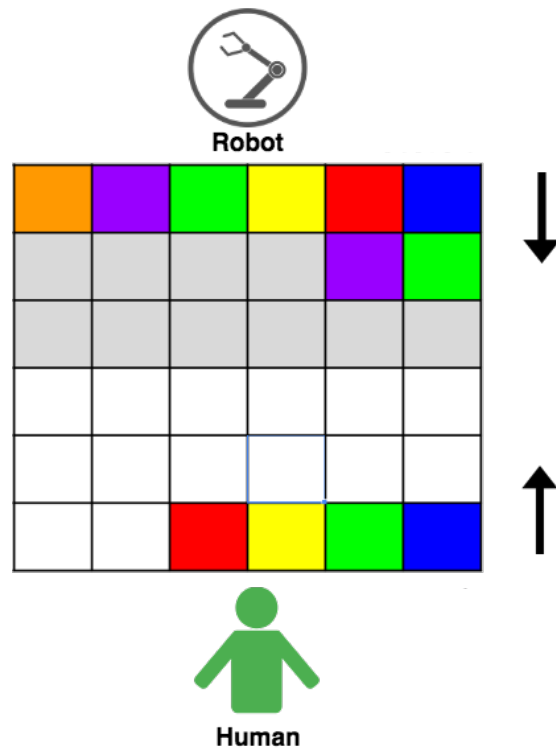
Shaping **Humans** to Match **Robot** Expectations!

# User Study



# Realtime Color Sudoku:

A really hard game for humans

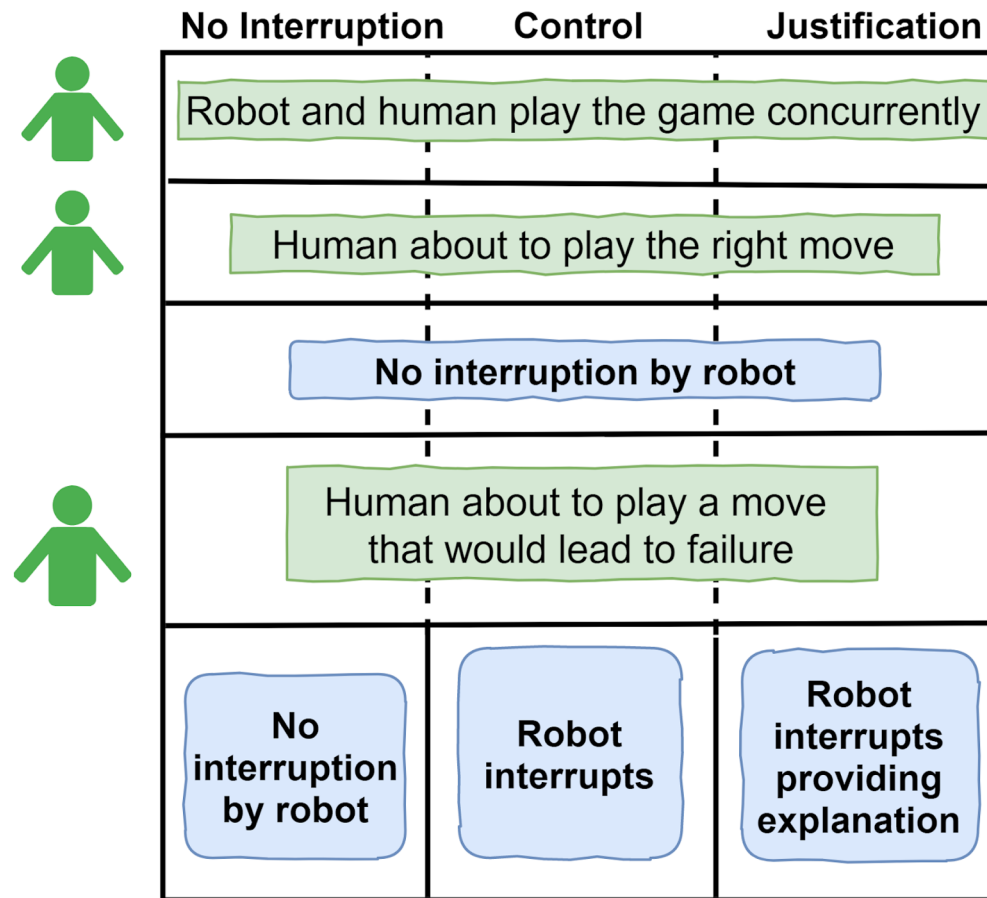


Each player gets 3 rows to fill:  
near to far, right to left.

There are no turns:  
play whenever you're ready



# Between-subjects experiment (n=26)



## Control:

Players about to make a mistake were told that they cannot make that move or they'll fail the game.

## Justification:

Players about to make a mistake were told about the reward inferred they were missing.

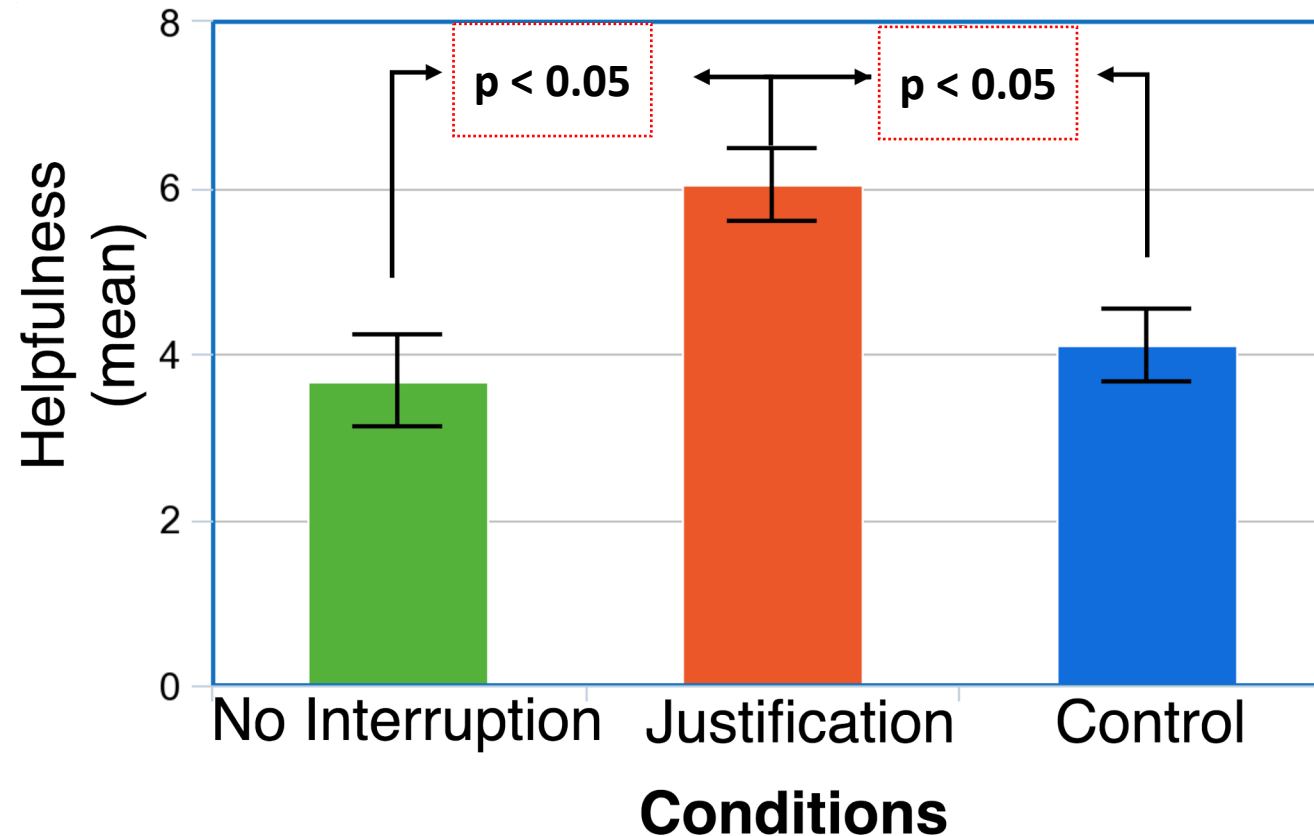
## No Interruption:

Players completed the game without mistakes.

# Subjective Hypotheses

- H1: Participants will find the robot more helpful and useful when it explains why a failure may occur
- H2: Participants will find the robot to be more intelligent when providing justification for its advice

# Subjective Results: Helpfulness



H1: **Participants will find the robot more helpful and useful when it explains why a failure may occur**



# Subjective Hypotheses

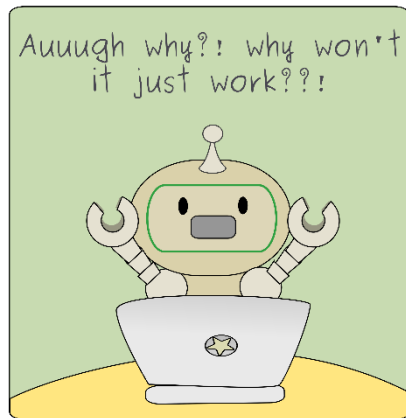
- ✓ H1: Participants will find the robot more helpful and useful when it explains why a failure may occur
- ✓ H2: Participants will find the robot to be more intelligent when providing justification for its advice

# Objective Hypothesis

H1: Participants will complete the game faster when provided with justification

But we couldn't test it.

Because most participants *didn't even listen* to the control condition's advice without justification.



Game Completion Rate:

Control: 20%

Justification: 80%



# Explainable AI for Establishing Shared Expectations During Human-Robot Collaboration

Collaborative Artificial Intelligence and Robotics Lab



Prof. Brad Hayes

Bradley.Hayes@Colorado.edu

<http://www.cairo-lab.com/>

<http://www.circadence.com/>

 @hayesbh

 <http://bradhayes.info>