Symbols as a Lingua Franca For Supporting Human-AI Interaction For Explainable and Advisable AI Systems

(Human-AI Interaction beyond Nodding & Pointing)

Subbarao Kambhampati
School of Computing & AI
Arizona State University

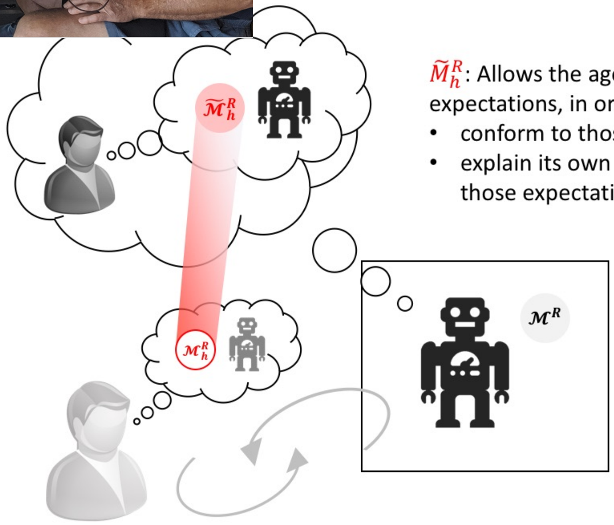
Research Funded in part by

Email: rao@asu.edu Twitter: @rao2z Linkedin: @Subbarao2z

Seminar @ AI Institute for Advances in Optimization

Research Background..

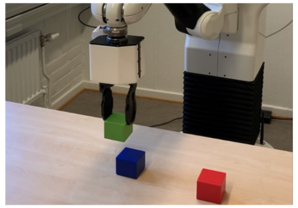
- We have focused on explainable human-AI interaction.
- Our setting involves collaborative problem solving, where the AI agents provide decision support to the human users in the context of *explicit knowledge sequential decision-making tasks* (such as mission planning)
  - In contrast, much work in social robotics and HRI has focused on tacit knowledge tasks (thus making explanations mostly moot)
  - We assume that the AI agent either learns the human model or has prior access to it.
- We have developed frameworks for proactive explanations based on *model reconciliation* as well as on-demand *foil-based explanations*
- We have demonstrated the effectiveness of our techniques with systematic (IRB approved) human subject studies

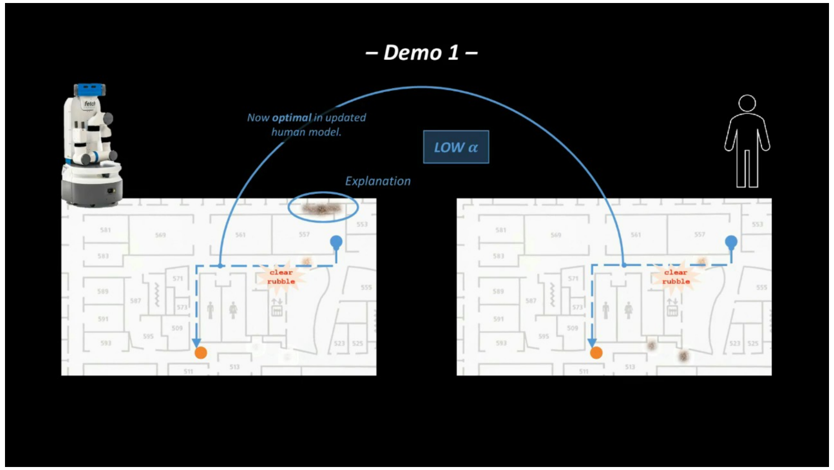


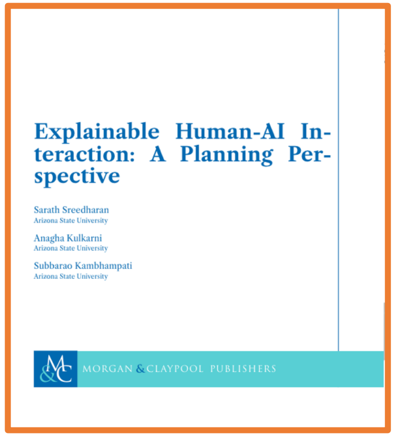
$\widetilde{M}_h^R$: Allows the agent to anticipate human expectations, in order to
- conform to those expectations
- explain its own behavior in terms of those expectations.

M_r^H and $\widetilde{M}_h^R$ are *Expectations* on Models $\mathcal{M}^H$ and $\mathcal{M}^R$

They don't have to be executable







Explainable Human-AI Interaction: A Planning Perspective

Sarath Sreedharan
Arizona State University

Anagha Kulkarni
Arizona State University

Subbarao Kambhampati
Arizona State University

MORGAN & CLAYPOOL PUBLISHERS

Viewpoint

Polanyi's Revenge and AI's New Romance with Tacit Knowledge

Artificial intelligence systems need the wisdom to know when to take advice from us and when to learn from data.

IN HIS 2019 Turing Award Lecture, Geoff Hinton talks about two approaches to make computers intelligent. One he dubs—tongue firmly in cheek—"Intelligent Design" (or giving task-specific knowledge to the computers) and the other, his favored one, "Learning" where we only provide examples to the computers and let them learn. Hinton's not-so-subtle message is that the "deep learning revolution" shows the only true way is the second.

Hinton is of course reinforcing the AI zeitgeist, if only in a doctrinal form. Artificial intelligence technology has captured popular imagination of late, thanks in large part to the impressive feats in perceptual intelligence—including learning to recognize images, voice, and rudimentary language—and bringing fruits of those advances to everyone via their smartphones and personal digital accessories. Most of these advances did indeed come from "learning" approaches, but it is important to understand the advances have come in

"Human, grant me the serenity to accep
things I cannot learn, data to learn the t
I can, and wisdom to know the differenc

signed—for which we do have explicit

Twitter @rao2z

subbarao kambhampati

6 RESULTS

SORT BY: NEWEST OLDEST RELEVANCE

What just happened? The rise of interest in artificial intelligence

TECHNOLOGY — 08/11/2019

145

Perception won't be reality, once AI can manipulate what we see

DOI:10.1145/3546954 https://cacm.acm.org/blogs/blog-cacm

Changing the Nature of AI Research

Subbarao Kambhampati considers how artificial intelligence may be straying from its roots.

Subbarao Kambhampati
AI as (an Ersatz) Natural Science?
https://bit.ly/3Rcf5NW
June 8, 2022

In many ways, we are living in quite a wondrous time for artificial intelligence (AI), with every week bringing some awe-inspiring feat in yet another tacit knowledge (https://bit.ly/3qYrAOY) task that we were sure would be out of reach of computers for quite some time to come. Of particular recent interest are the large learned systems based on transformer architectures that are trained with billions of parameters over massive Web-scale multimodal corpora. Prominent examples include large language models (https://bit.ly/3iGdekA) like GPT3 and PALM that respond to free-form text prompts, and language/image models like DALL-E and Imagen that can map text prompts to photorealistic images tal ways. Just the other day, some researchers were playing with DALL-E and thought that it seems to have developed a secret language of its own (https://bit.ly/3ahH1Py) which, if we can master, might allow us to interact with it better. Other researchers found that GPT3's responses to reasoning questions can be improved by adding certain seemingly magical incantations to the prompt (https://bit.ly/3aelxmI), the most prominent of these being "Let's think step by step." It is almost as if the large learned models like GPT3 and DALL-E are alien organisms whose behavior we are trying to decipher.

This is certainly a strange turn of events for AI. Since its inception, AI has existed in the no-man's land between engineering (which aims at designing systems for specific functions), and "Science" (which aims to discover the regularities in naturally occurring phenomena). The science part of AI came from its original pre- havior) rather than on insights about natural intelligence.

This situation is changing rapidly–especially as AI is becoming synonymous with large learned models. Some of these systems are coming to a point where we not only do not know how the models we trained are able to show specific capabilities, we are very much in the dark even about what capabilities they might have (PALM's alleged capability of "explaining jokes"—https://bit.ly/3yJk1m4— is a case in point). Often, even their creators are caught off guard by things these systems seem capable of doing. Indeed, probing these systems to get a sense of the scope of their "emergent behaviors" has become quite a trend in AI research of late.

Given this state of affairs, it is increasingly clear that at least part of AI is straying firmly away from its "engineering" roots. It is increasingly hard to consider large learned systems as "designed" in the traditional sense of the

getting computers to learn to do tasks
for which we only have tacit knowl-

Beauty, lies & ChatGPT: Welcome to the post-truth world

BY SUBBARAO KAMBHAMPATI, OPINION CONTRIBUTOR - 02/16/23 10:00 AM ET

SHARE TWEET

AP Photo/Timothy D. Easley

Bella Whitice talks with classmate Katherine McCormick as they try and outwit the "robot" that was creating writing assignments in Donnie Piercey's class at Stonewall Elementary in Lexington, Ky., Monday, Feb. 6, 2023. The robot was the new artificial intelligence tool ChatGPT which can generate everything from essays and haikus to term papers in a matter of seconds.

Two months back, a company called OpenAI released its chatbot, ChatGPT, to the public. ChatGPT is a so-called Large Language Model (LLM) that is

Just In...

The Hill's Morning Report — Political focus on Ohio derailment intensifies
MORNING REPORT - 26M 30S AGO

Space telescope uncovers massive galaxies near cosmic dawn
NEWS - 50M 57S AGO

Railroad pushback to safety regulations scrutinized amid East Palestine disaster
LOBBYING - 56M 30S AGO

How the Russia-Ukraine war has changed cyberspace
CYBERSECURITY - 56M 30S AGO

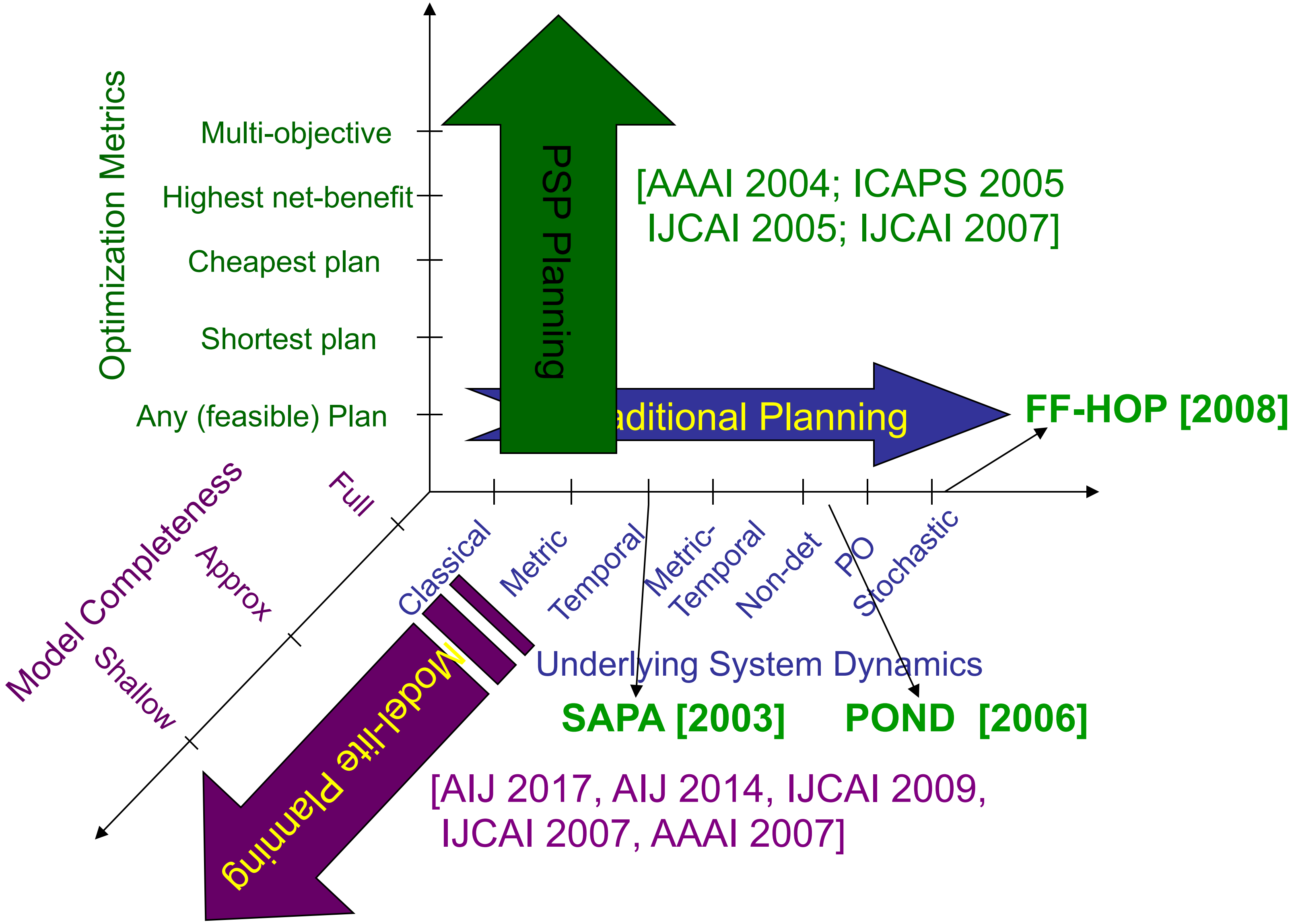
What to expect in the second year of the Russia-Ukraine war
DEFENSE - 56M 30S AGO

Greene stirs up political storm with 'national divorce' comments

Most Popular

1 Most young men are single. Most young women are not.

2 Biden declines to veto Apple Watch ban


Optimization Metrics
Multi-objective
Highest net-benefit
Cheapest plan
Shortest plan
Any (feasible) Plan
PSP Planning
[AAAI 2004; ICAPS 2005
IJCAI 2005; IJCAI 2007]
aditional Planning
FF-HOP [2008]
Model Completeness
Full
Approx
Shallow
Classical
Metric
Temporal
Metric-
Temporal
Non-det
PO
Stochastic
Underlying System Dynamics
SAPA [2003]
POND [2006]
Model-lite Planning
[AIJ 2017, AIJ 2014, IJCAI 2009,
IJCAI 2007, AAAI 2007]

Planning continues to be a shibb

- Even though LLM's can slice and dice explain jokes, they still can't plan!
  - New Captcha?
- Of course, they can be used as heuristic guidance for an underlying sound planner
  - But hey, pretty much anything can be a "heuristic"..
- Grammar vs Meaning..

arXiv > cs > arXiv:2206.10498

Computer Science > Computation and Langu

[Submitted on 21 Jun 2022]

Large Language Models Still Ca Planning and Reasoning about

Karthik Valmeekam, Alberto Olmo, Sarath Sreed

The recent advances in large language models (LLMs From GPT-3 to PaLM, the state-of-the-art performan new large language model. Along with natural langu whether such model interest in developir benchmarks seem n these benchmarks c reasoning capabiliti

```
[STATEMENT]
As initial conditions I hav
, the yellow block is cl
orange block, the red
and the yellow block is
My goal is to have that the
```

NEURAL AI & futurism powered by IBM

This article was published on July 31, 2022

ARTIFICIAL INTELLIGENCE

Large language models can't plan, even if they write fancy essays

Large language models perform very poorly at tasks that require methodical planning

July 31, 2022 - 8:50 pm

Popular on Neural

1. Synthetic data is the safe, low-cost alternative to real data that we need
2. A new 'common sense' test for AI could lead to smarter machines
3. Climate risks are a major business threat – here's how AI can help
4. Confused Replika AI users are standing up for bots and trying to bang the algorithm
5. New experiment demonstrates that reality might actually be real

This article is part of our coverage of the latest in AI research.

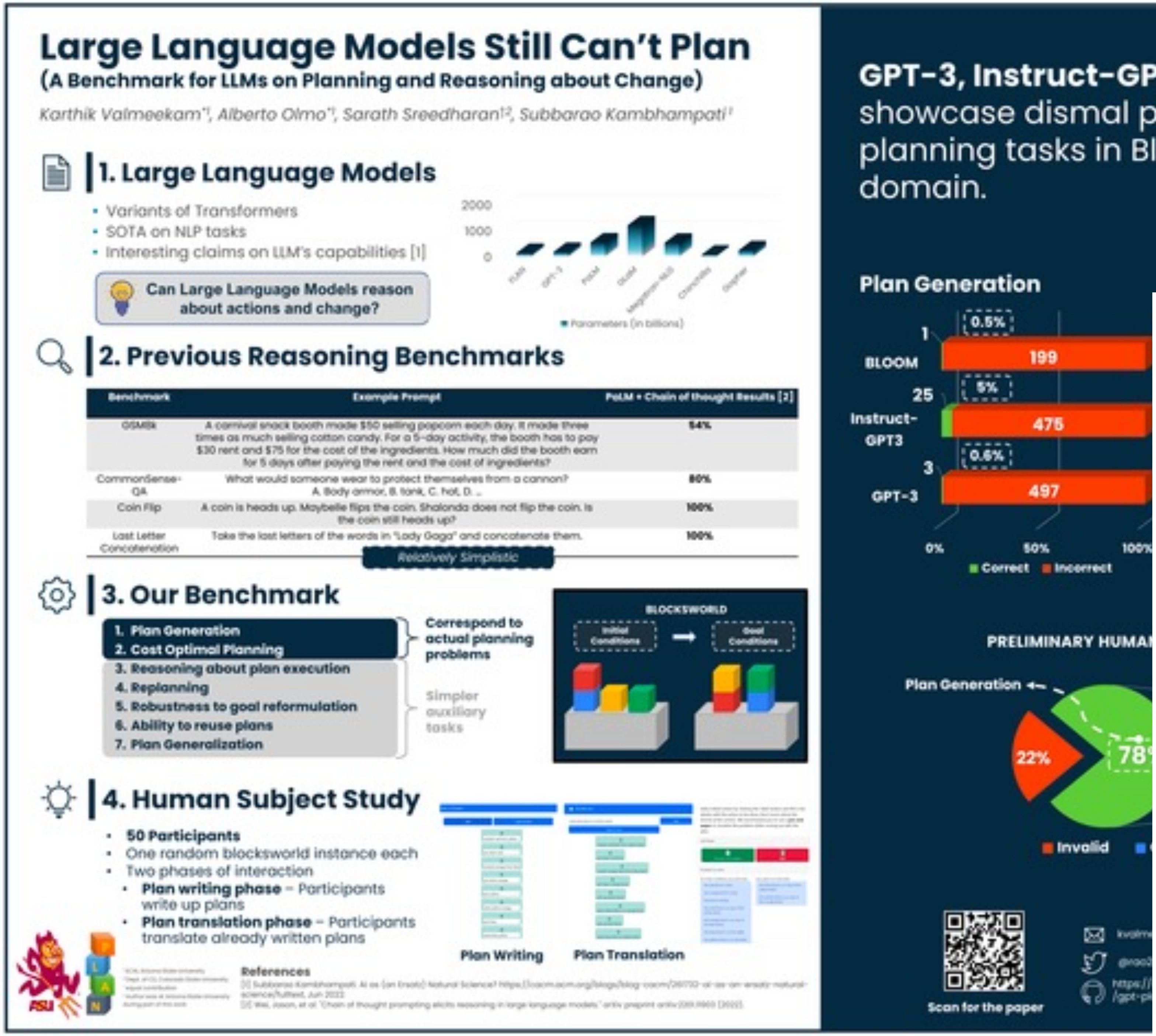
Plan Reuse We showcase an instance and the respective pl chine with an instance which requires only a certain prefix of the plan provided in the example	0/500 = 0%
Plan Generalization We showcase an instance and the respective plan as an example and prompt the machine with a new instance. The plans for both the instances can be generated by a fixed program containing loops and conditionals.	33/500 = 6.6%
Goal Directed Reasoning We showcase an instance and the respective plan as an example and prompt the machine with a new instance.	3/500 = 0.6%
Optimal Planning We showcase an instance, the respective optimal plan and the associated cost as an example and prompt the machine with a new instance.	1/500 = 0.2%
Replanning We showcase an instance, the respective plan and present an unexpected change of the state. We then also present a new plan from the changed state. Finally, for a new instance we repeat the same except we ask the machine for the new plan.	0/500 = 0%

Table 1: LLM Assessment Suite Results on Davinci (base version)

```
(pick-up orange)
(stack orange red)

==================================SUCCESS==================================
```

Listing 2: Goal directed reasoning



Large language models can't plan, even if they write fancy essays

Large language models perform very poorly at tasks that require methodical planning

July 31, 2022 - 8:50 pm

Popular on Neural

1 Concrete is one of the world's most harmful materials. Graphene could change that

2 AI translation firm unveils 'world-first' timeline to singularity

3 Confused Replika AI users are standing up for bots and trying to bang the algorithm

4 What to expect from AI in 2023

5 Quantum computing startup eyes mainstream adoption after £30m investment

STORY BY
Ben Dickson

This article is part of our coverage of the latest in AI research.

Large language models like GPT-3 have advanced to the point that it has become difficult to measure the limits of their capabilities. When you have a very large neural network that can generate articles, write software code, and engage in conversations about sentience and life, you should expect it to be able to reason about tasks and plan as a human does, right?

Wrong. A study by researchers at Arizona State University, Tempe, shows that when it comes to planning and thinking methodically, LLMs perform very poorly, and suffer from many of the same failures observed in current deep learning systems.

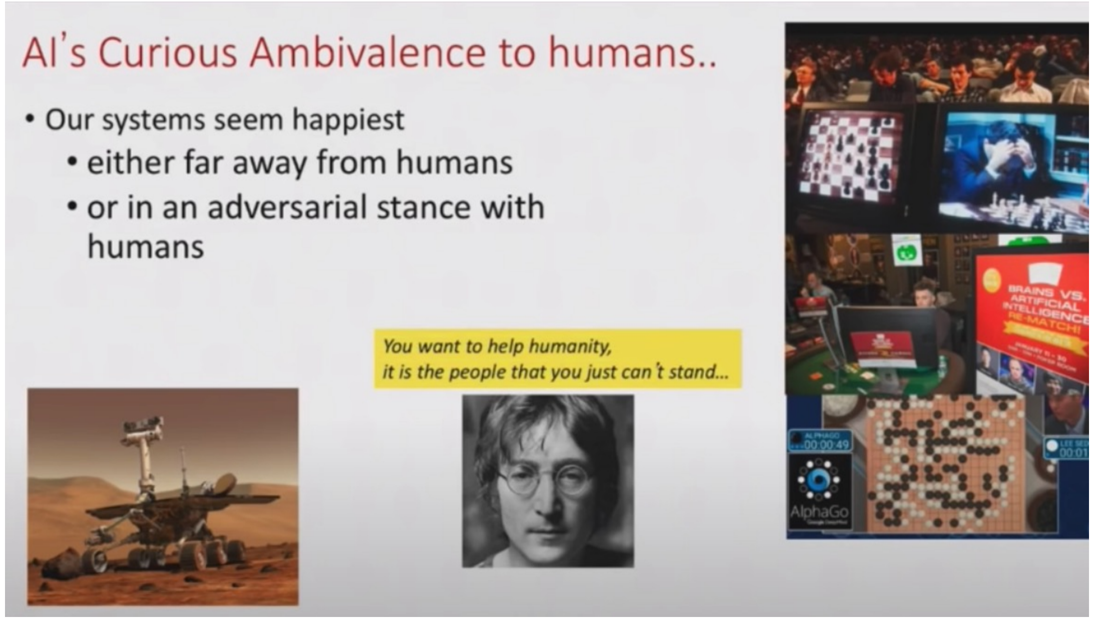

My other agent is a $&^&%
HUMAN
Other Agents
~~God~~

(Ignoring) Humans: AI vs OR

shutterstock.com · 1195798915

The AI Way: If you stay far enough away from 'em (e.g. Mars) or in adversarial stance with 'em (e.g. Zero-Sum games), you'll be fine..

- or drag them into the land of AI..



The OR Way: We will send one of our guys (the hORse whisperer) along with our methods—and the guy will do the "human interaction"

- ..and train *our guys* to learn to deal with humans

DECISION ANALYSIS
Vol. 9, No. 3, September 2012, pp. 274–292
ISSN 1545-8490 (print) | ISSN 1545-8504 (online)
http://dx.doi.org/10.1287/deca.1120.0238

Communicating Analytic Results: A Tutorial for Decision Consultants

Jeffrey M. Keisler
Management Science and Information Systems Department, College of Management, University of Massachusetts Boston, Boston, Massachusetts 02125, jeff.keisler@umb.edu

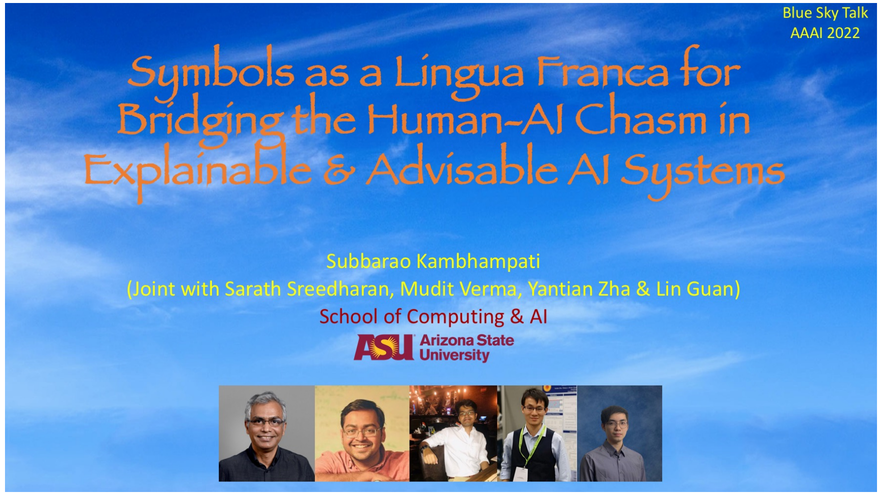
Patrick S. Noonan
Goizueta Business School, Emory University, Atlanta, Georgia 30322, patrick.noonan@bus.emory.edu

Good analysis alone may not achieve the goals of decision analysis (DA) engagements. Good communication of the results of that analysis can help stakeholders understand, accept, and implement the recommended course of action. Practitioners can use decision-analytic principles when considering the decision of how to communicate results themselves. From this perspective, we consider a range of questions to ask in preparing for communication with the client and other stakeholders. We review standard communication practices in DA engagements. The standard practice can be improved by drawing on insights from other areas of management practice. Decision analysis has both technical and organizational features, and we discuss ways to deal with the conceptual and expressive challenges this presents. This pragmatic tutorial provides a starting point for decision analysts to develop both technical communication skills and organizational communication skills.

Key words: communication of decision analysis insights; modeling; analysis; decision analysis; decision consulting
History: Received on December 28, 2011. Accepted on February 14, 2012, after 1 revision. Published online in *Articles in Advance* August 3, 2012.

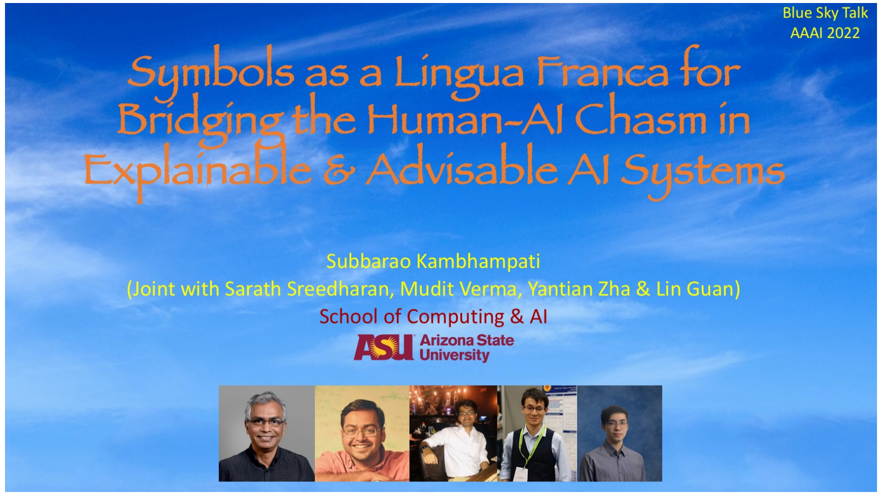
Talk Overview

- Part 1: Why and how do humans exchange explanations? Do AI systems need to?
- Part 2: Using Mental Models for Explainable Behavior in the context of explicit knowledge tasks (think Task Planning)
  - The 3-model framework: $\mathcal{M}^R, \mathcal{M}^H, \mathcal{M}_h^R$
  - Explicability: *Conform to* $\mathcal{M}_h^R$
  - Explanation: Reconcile $\mathcal{M}_h^R$ to $\mathcal{M}^R$
  - Extensions: *Foils, Abstractions, Multiple Humans..*
- Part 3: Supporting explainable behavior even without shared vocabulary
  - Symbols as a *Lingua Franca* for Explainable and Advisable Human-AI Interaction
    - *Post hoc symbolic explanations of inscrutable reasoning*
    - *Accommodating symbolic advice into inscrutable systems*



Talk Overview

- Part 1: Why and how do humans exchange explanations? Do AI systems need to?
- Part 2: Using Mental Models for Explainable Behavior in the context of explicit knowledge tasks (think Task Planning)
  - The 3-model framework: $\mathcal{M}^R, \mathcal{M}^H, \mathcal{M}_h^R$
  - Explicability: *Conform to* $\mathcal{M}_h^R$
  - Explanation: Reconcile $\mathcal{M}_h^R$ to $\mathcal{M}^R$
  - Extensions: *Foils, Abstractions, Multiple Humans..*
- Part 3: Supporting explainable behavior even without shared vocabulary
  - Symbols as a *Lingua Franca* for Explainable and Advisable Human-AI Interaction
    - *Post hoc symbolic explanations of inscrutable reasoning*
    - *Accommodating symbolic advice into inscrutable systems*



How do Humans Exchange Explanations?

- **Pointing (Tacit) Explanations**
    - Pointing to specific features of the object/image etc.
  - Feasible sometimes for one-shot classification decisions on spatial data (point to the right parts of the image/object)
    - "This is is a Red Striped Butterfly because…(Show)"
  - But quite unwieldy ["High Band Width AND Cognitive Load"] for explaining sequential decisions on spatio/temporal data (as it will involve pointing to the relevant regions of the space-time tube..)
    - "The reason I took this earlier United Flight is because… (point to the video of your life?)"
- **Symbolic (Explicit) Explanations**
  - Feasible for both spatial and spatio-temporal data and one-shot or sequential decisions
  - Requires that the humans share a symbolic vocabulary   (..or learn one to get by..)

- Typically, pointing explanations are used for tacit knowledge tasks, and symbolic ones for explicit knowledge tasks.
  - However, over time, we tend to develop symbolic vocabulary for exchanging explanations even for tacit knowledge tasks.
    - Consider, for example, Pick-and-Roll in Basketball..
- Symbolic explanations are not just "compact" but significantly reduce cognitive load on the receiver
  - (even though the receiver likely has to re-create the space-time tube versions of those explanations within their own minds)

But (Why) Do AI Systems have to give Explanations?

- Internal (Self) explanations within the system
    - "Soliloquy"
  - Explanations (e.g. "nogoods") to guide search
  - Explanations to guide learning: EBL
- External Explanations
  - To other systems
    - (offering proofs of correctness of decisions)
  - To the humans in the loop
    - Can't be a "Soliloquy"—unless the humans have no life but to understand the system's mutterings..
    - Explanation depends on the role of the human
      - "Debugger": Humans who are willing to go into the land of the machine just to figure out what it is doing
      - "End User"—Observer/Collaborator/Student/Teacher: Want rationales for the machine decisions that are comprehensible to them (without having to read huge manuals)
- (XAI has typically been about Explanations to Humans in the loop—but is often confused with techniques more relevant to the other settings)

Facebook makes millions of recommendations per day, and no one asks for an explanation!

--A Facebook AI Bigwig

Use cases for Human-In-The-Loop Explanations

- Debugger trying to flag and correct the system's behavior
- Observer (Lay)
- Observer (Expert)
- Collaborator (on a joint task)
- Student (Machine is in a teaching role)
- Teacher (Machine is in a learning role)

- Task: One shot vs. Sequential decision
- Interaction: One epoch vs. longitudinal

Explanations: Given to a specific human

Interpretability: Can humans make sense of it?

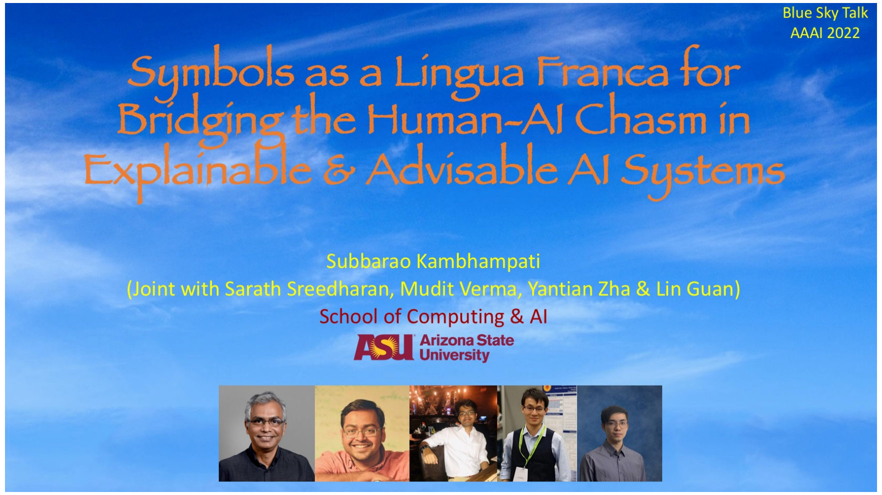
Certificates of Correctness: Given to a human's AI

Requirements on Explanations

- Comprehensibility
  - Cognitive load in parsing the explanation [Is the explanation in a form/level that is accessible to the receiving party]
- Communicability
  - Ease of exchanging the explanation
- Soundness
  - A guarantee from the other party that this explanation is really the reason for the decision
  - Related: Guarantee (to stand behind the explanation)
    - We expect the decision to change when the explanation is falsified
- Satisfaction (with the explanation)
  - Unfortunately, this is a slippery slope. ”Sweet Little Lies” start right here..
    - Very important not to do an “end to end” learning on ”what explanations seem to make people happy”!
    - GDPR and GPT3!

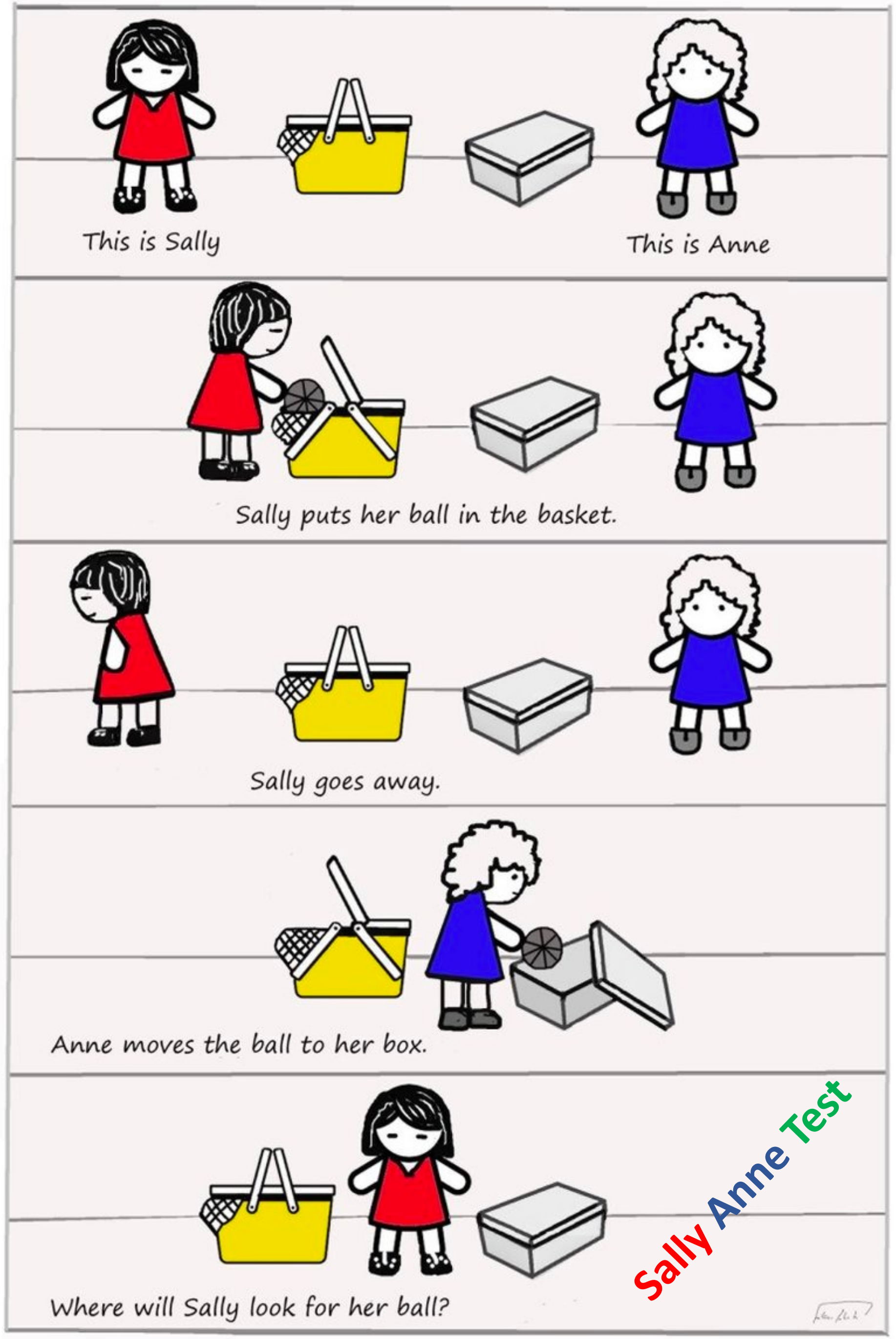
Talk Overview

- Part 1: Why and how do humans exchange explanations? Do AI systems need to?
- Part 2: Using Mental Models for Explainable Behavior in the context of explicit knowledge tasks (think Task Planning)
  - The 3-model framework: $\mathcal{M}^R, \mathcal{M}^H, \mathcal{M}_h^R$
  - Explicability: *Conform to* $\mathcal{M}_h^R$
  - Explanation: Reconcile $\mathcal{M}_h^R$ to $\mathcal{M}^R$
  - Extensions: *Foils, Abstractions, Multiple Humans..*
- Part 3: Supporting explainable behavior even without shared vocabulary
  - Symbols as a *Lingua Franca* for Explainable and Advisable Human-AI Interaction
    - *Post hoc symbolic explanations of inscrutable reasoning*
    - *Accommodating symbolic advice into inscrutable systems*



What does it take for an AI agent to show explainable behavior in the presence of human agents?

Managing Mental Models



Let’s start with the tale of three models..

```
(:action move
:parameters (?from ?to - location)
:precondition (and (robot-at ?from)
                   (hand-tucked)
                  (crouched))
:effect (and (robot-at ?to)
             (not (robot-at ?from))))

(:action tuck
:parameters ()
:precondition ()
:effect (and (hand-tucked)
            (crouched)     ))

(:action crouch
:parameters ()
:precondition ()
:effect (and (crouched)))
```

We will think of Models as

$$< I, G, A, O, \pi >$$

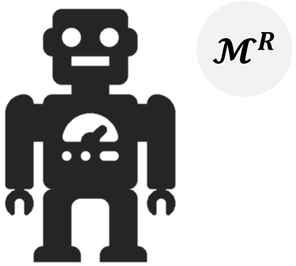
- I Initial state
- G Goals
- A Actions
- O Observation model
- π Plan

The development is largely agnostic to the specific framework

→ Relational representations PDDL
→ Dynamic Programming Rep MDP/RL

Start State $S \longrightarrow a_1 \longrightarrow a_2 \longrightarrow \circ \;\circ\; \circ \longrightarrow a_n \longrightarrow G$ **Goal State**

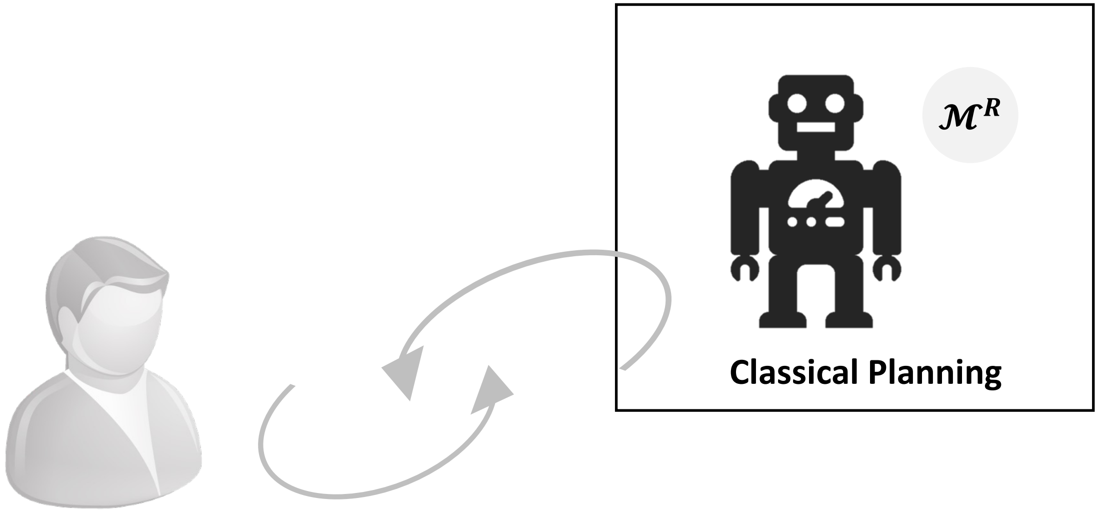
Given – $\color{red}{S}$, $\color{red}{G}$ and set of actions $\color{red}{\{a_i\}}$ => Agent's Model $\boldsymbol{M^R}$
*Find – sequence of actions or **plan** $\color{red}{\pi = \langle a_1, a_2, \dots, a_n \rangle}$ that transforms **S** to **G**.*



Classical Planning

Start State $\boxed{S} \rightarrow a_1^R \rightarrow a_2^H \rightarrow \circ \circ \circ \rightarrow a_n \rightarrow \boxed{G^+}$ **Goal State**

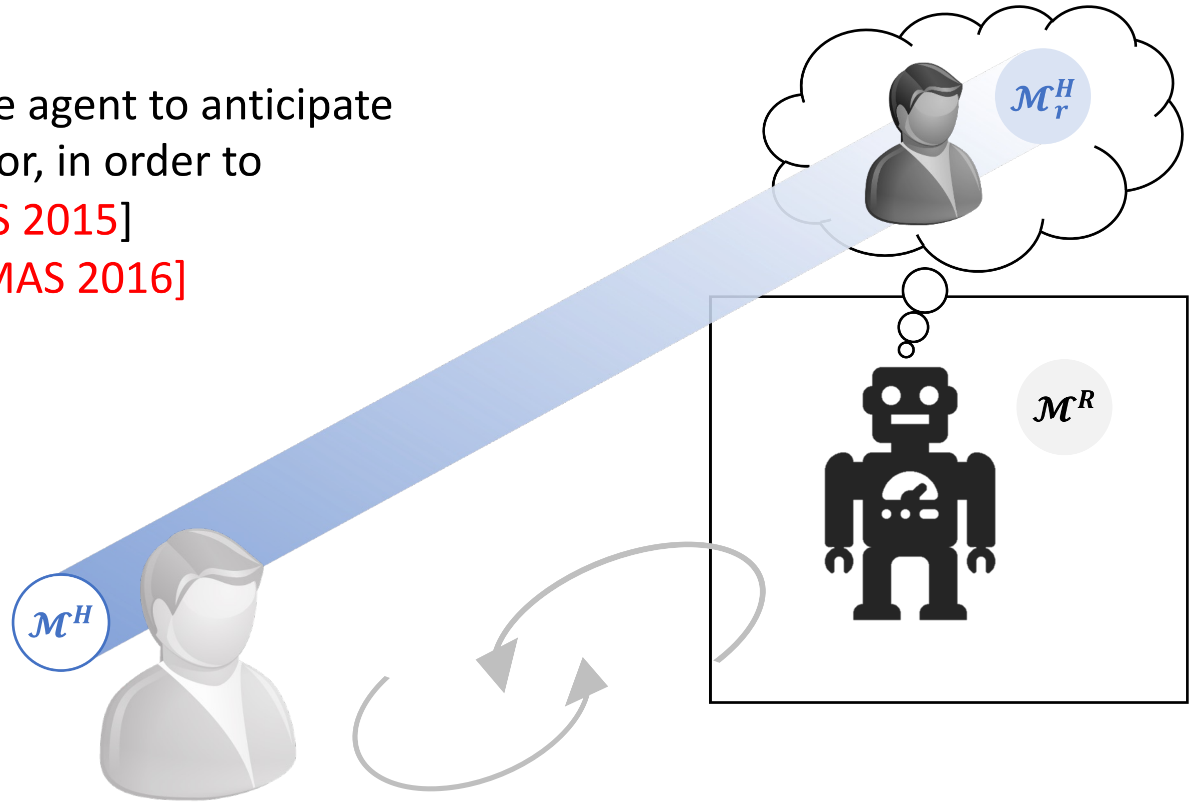
Given – S, G and set of actions $\{a_i\}$ *=> Agent's Model* $\boldsymbol{M}^R$
Find – sequence of actions or ***joint plan*** $\pi = \langle a_1, a_2, \ldots, a_n \rangle$ *that transforms* ***S*** *to* $\boldsymbol{G}^+$.



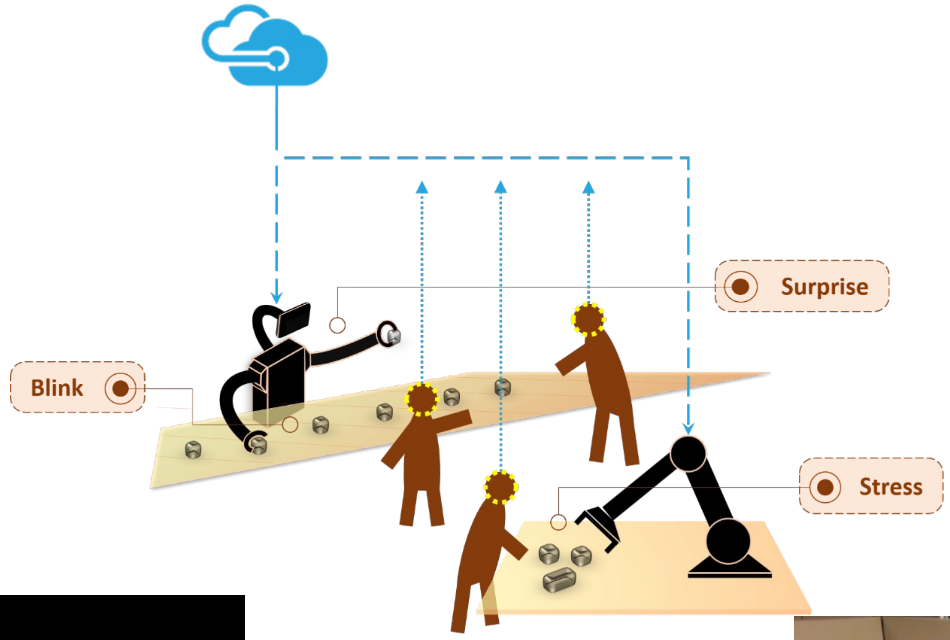
Start State $\boxed{S} \longrightarrow a_1^R \longrightarrow a_2^H \longrightarrow \circ \;\circ\; \circ \longrightarrow a_n \longrightarrow G^+$ **Goal State**

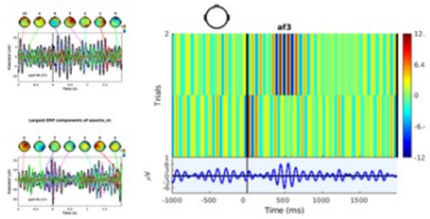
M_r^H: Allows the agent to anticipate human behavior, in order to

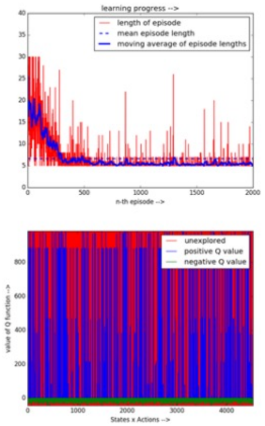
- assist [IROS 2015]
- avoid [AAMAS 2016]
- team, etc.



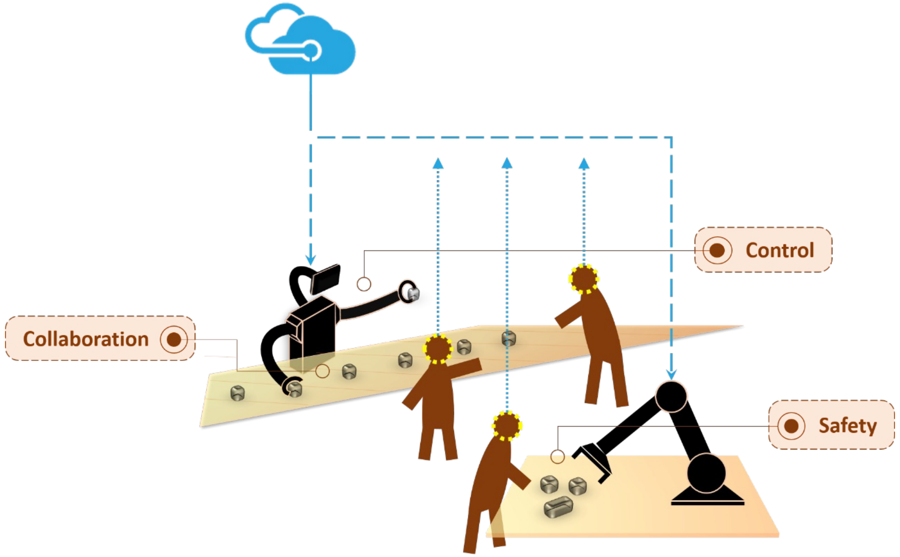
Intention Recognition with Emotive

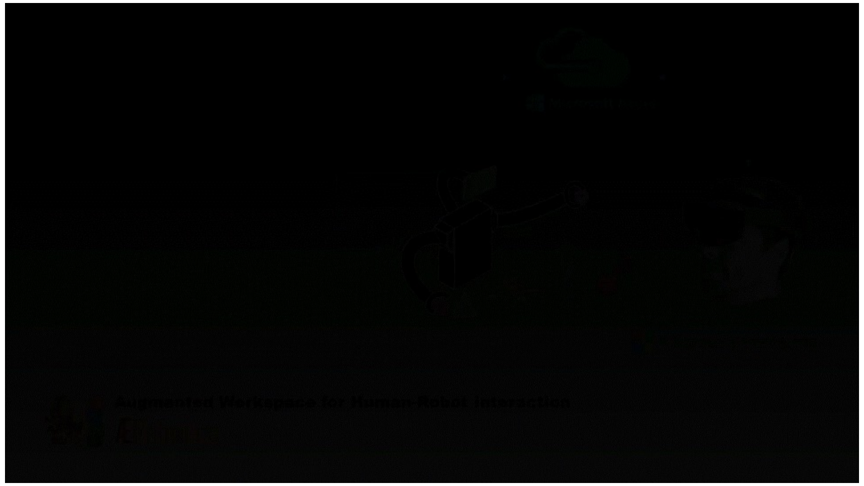






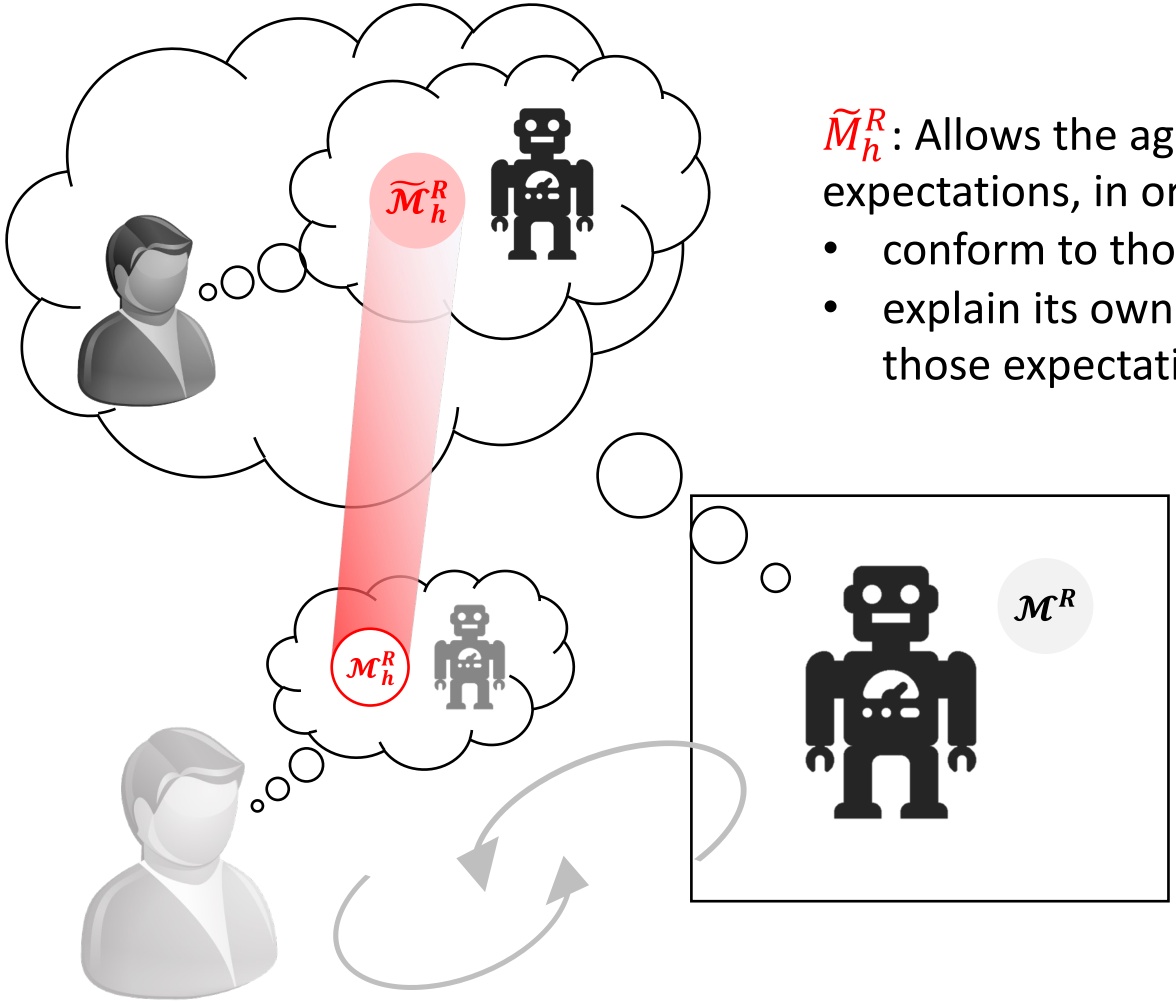
Intention Projection with Hololens





[IROS 2018]

Start State $S \rightarrow a_1^R \rightarrow a_2^H \rightarrow \circ \circ \circ \rightarrow a_n \rightarrow G^+$ Goal State



$\widetilde{M}_h^R$: Allows the agent to anticipate human expectations, in order to

- conform to those expectations
- explain its own behavior in terms of those expectations.

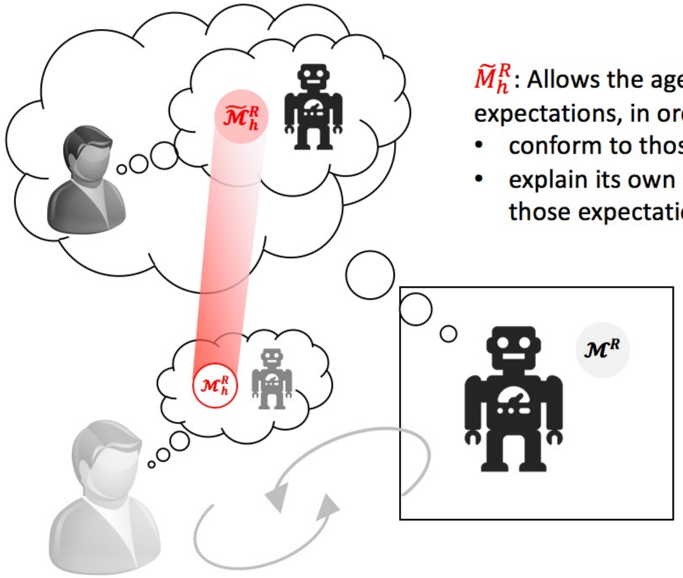
M_r^H and $\widetilde{M}_h^R$ are *Expectations* on Models $\mathcal{M}^H$ and $\mathcal{M}^R$

They don't have to be executable

Model differences with human in the loop

- The robot's task model may differ from the human's expectation of it
  - *Consequence →*
    - **Plans that are optimal to the robot may not be so in human's expectation**

      → ***"Inexplicable" plans***



$\widetilde{M}_h^R$: Allows the agent to anticipate human expectations, in order to

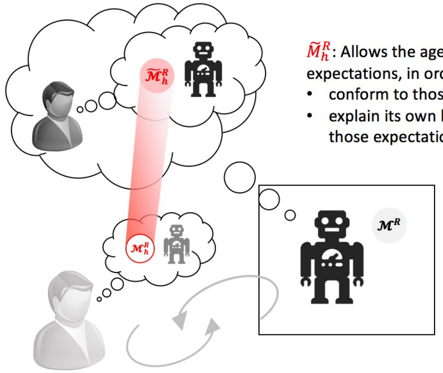
- conform to those expectations
- explain its own behavior in terms of those expectations.

Model differences with human in the loop

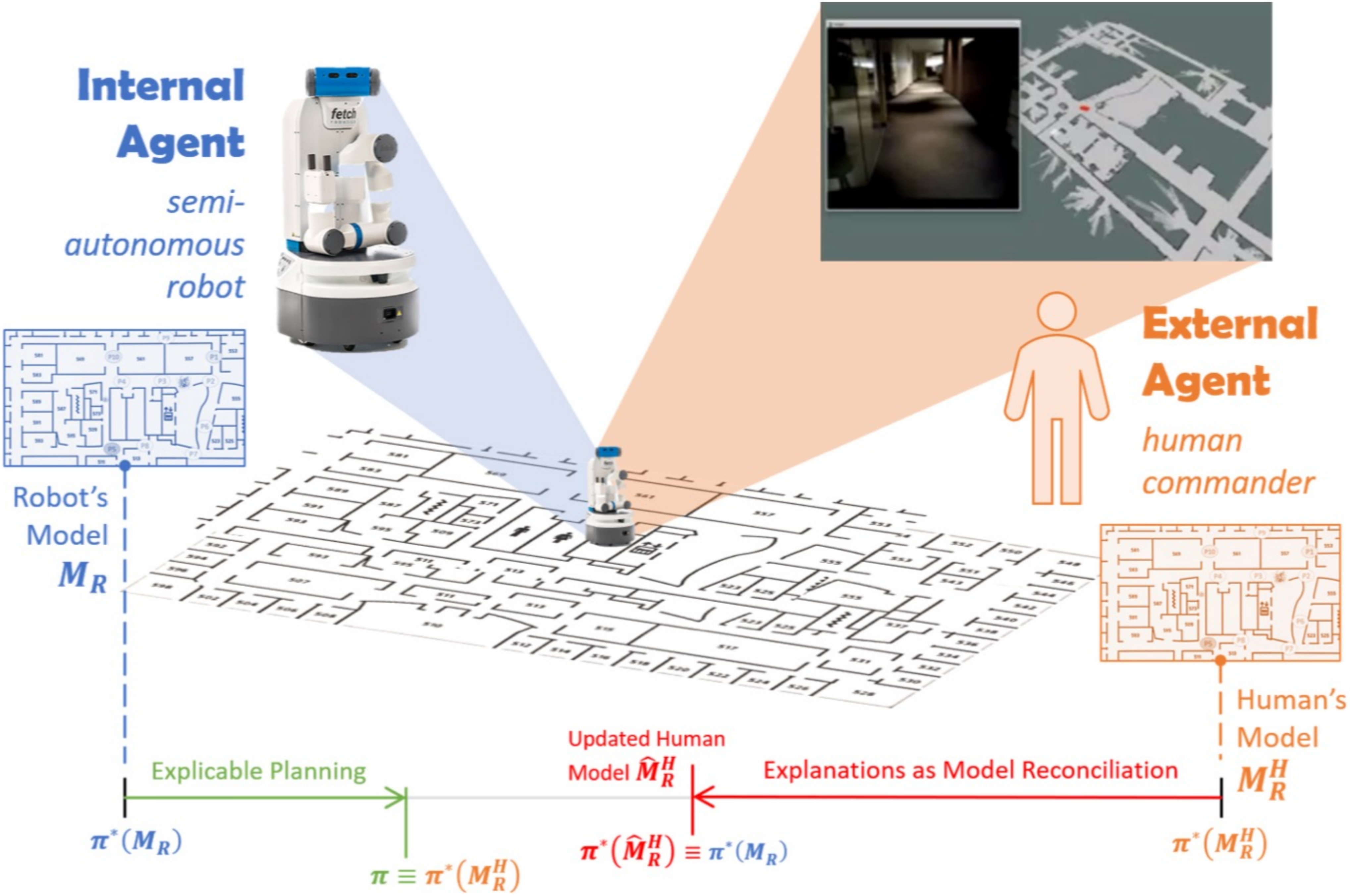
- The robot's task model may differ from the human's expectation of it
  - *Consequence* →
    - **Plans that are optimal to the robot may not be so in human's expectation**

      → ***"Inexplicable" plans***

- The robot then has **two options** –*conform to expectations or change them*
  - **Explicable planning –** sacrifice optimality in own model to be explicable to the human
  - **Plan Explanations –** resolve perceived suboptimality by revealing relevant model differences

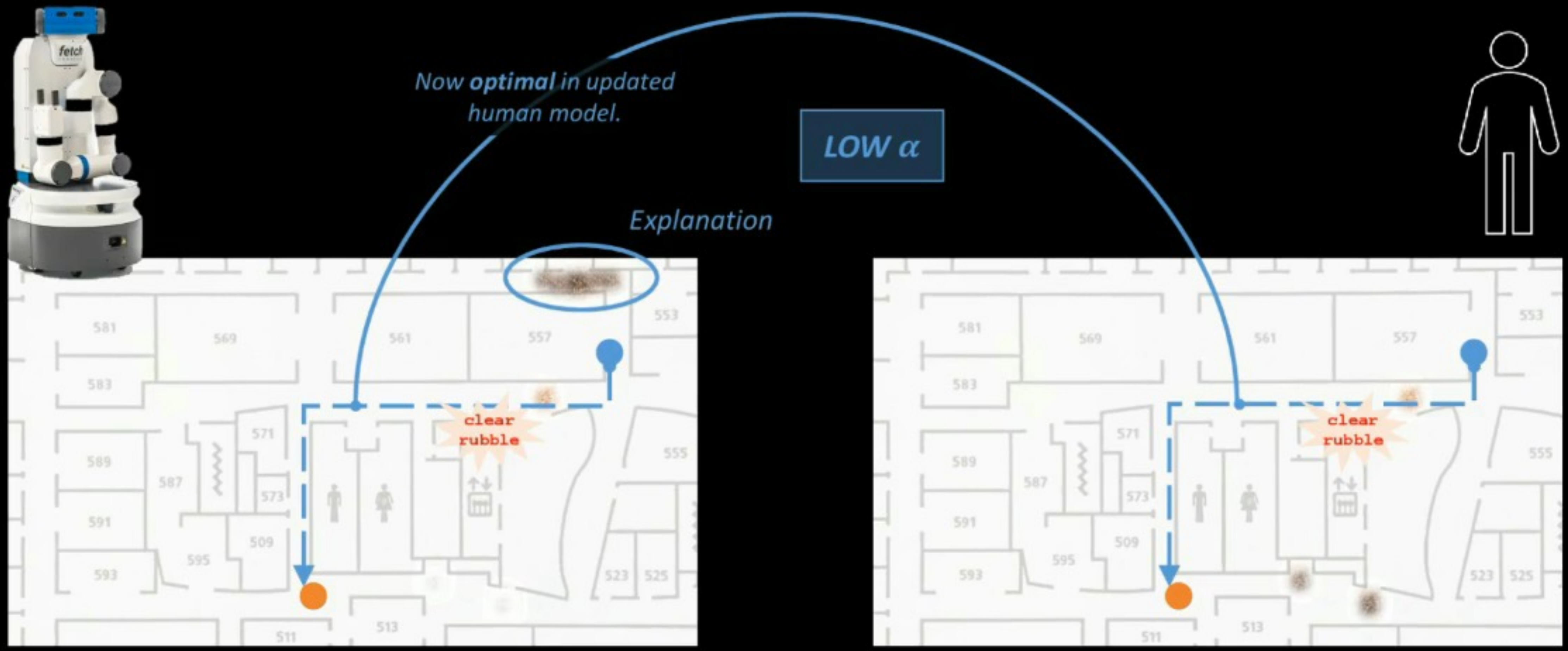


$\widetilde{M}_h^R$: Allows the agent to anticipate human expectations, in order to
- conform to those expectations
- explain its own behavior in terms of those expectations.



Internal Agent
semi-autonomous robot
Robot's Model M_R
External Agent
human commander
Human's Model M_R^H
Updated Human Model $\widehat{M}_R^H$
Explicable Planning
Explanations as Model Reconciliation
$\pi^*(M_R)$
$\pi \equiv \pi^*(M_R^H)$
$\pi^*(\widehat{M}_R^H) \equiv \pi^*(M_R)$
$\pi^*(M_R^H)$

– Demo 1 –


Now **optimal** in updated human model.
LOW α
Explanation
fetch
581
569
561
557
553
583
clear rubble
571
555
589
587
573
591
509
595
593
523
525
511
513
581
569
561
557
553
583
clear rubble
571
555
589
587
573
591
509
595
593
523
525
511
513

Explicable Plan

Given a goal, the objective is to find an explicable robot plan:

$$\underset{\pi_{M_R}}{\operatorname{argmin}} \boxed{cost(\pi_{M_R})} + \alpha \cdot \boxed{dist(\pi_{M_R}, \pi_{\mathcal{M}_R^*})}$$

Cost of robot plan

Distance between robot plan and human's expectation of robot plan

Problem: Conforming to expectations can be costly

Explanations as Model Reconciliation

A Human-Aware Planning (HAP) Problem is a tuple $\langle \boldsymbol{\mathcal{M}^R}, \boldsymbol{\mathcal{M}_h^R} \rangle$ where $\boldsymbol{\mathcal{M}^R} = \langle \boldsymbol{D^R}, \boldsymbol{I^R}, \boldsymbol{G^R} \rangle$ is the planner's model of the planning problem, and $\boldsymbol{\mathcal{M}_h^R} = \langle \boldsymbol{D_h^R}, \boldsymbol{I_h^R}, \boldsymbol{G_h^R} \rangle$ is the human's understanding of the same.

$C(\pi, \mathcal{M})$ is the cost of solution (plan) of model $\mathcal{M}$ and $C^*_{\mathcal{M}}$ is cost of the optimal plan.

Explanation ϵ for plan π →

(1) $\widehat{\mathcal{M}}_h^R \leftarrow \mathcal{M}_h^R + \epsilon$

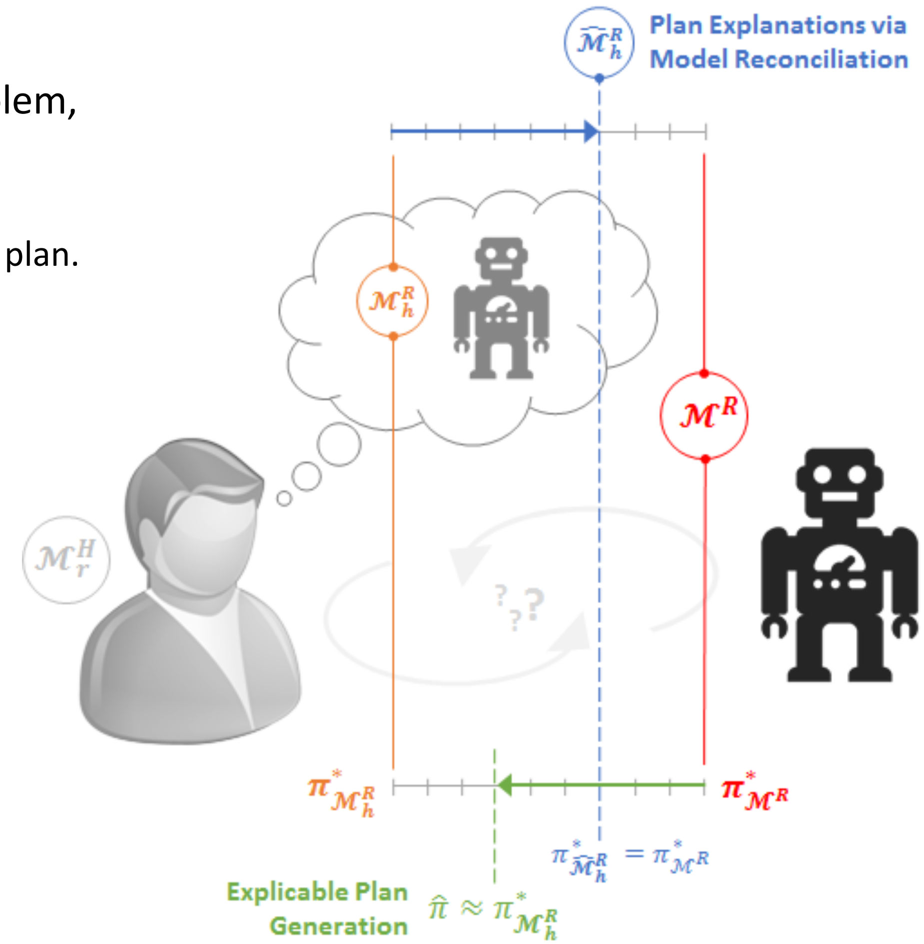
→ is a model update to the human

(2) $C(\pi, \mathcal{M}^R) = C^*_{\mathcal{M}^R}$

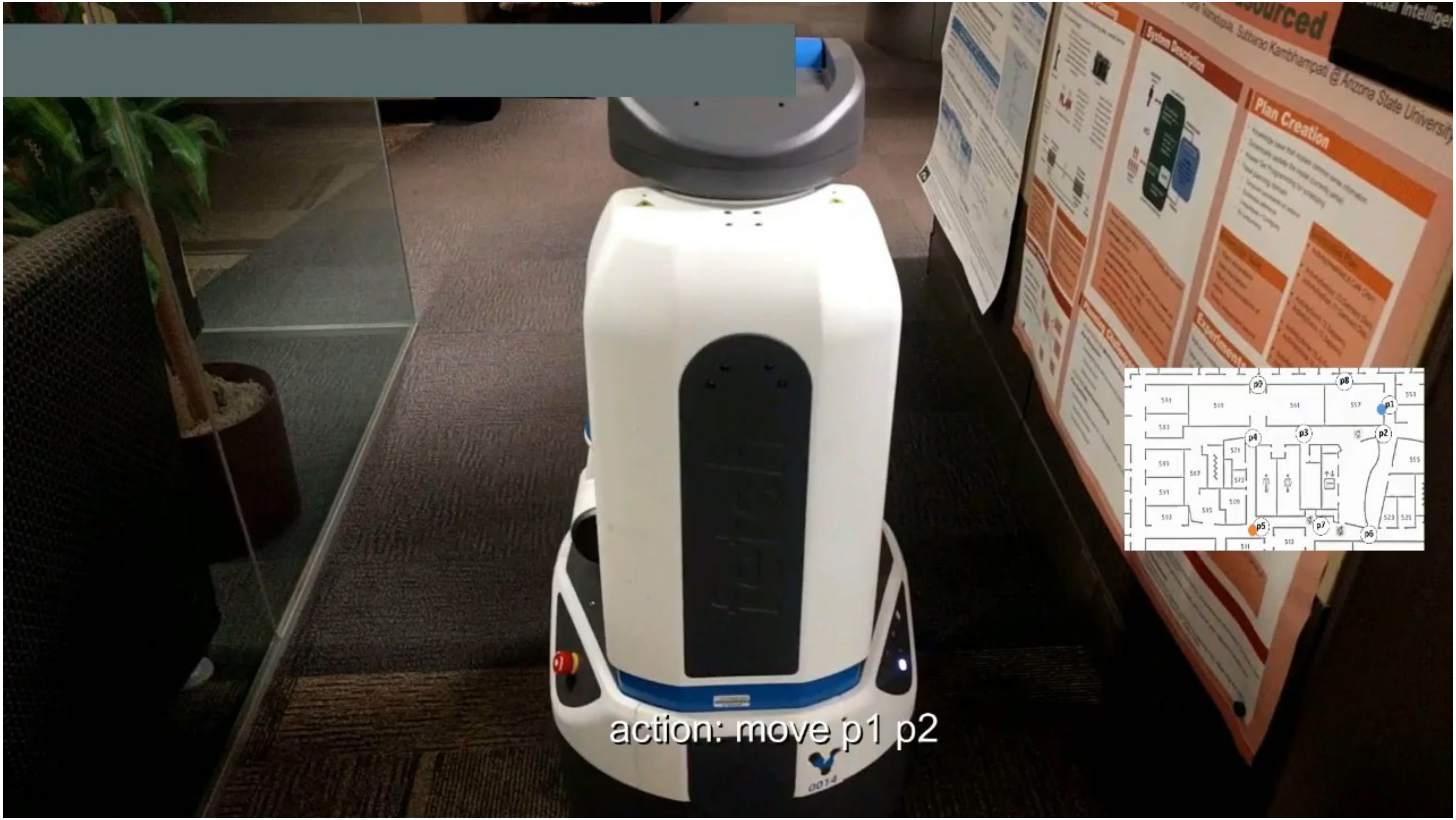
→ π is optimal in robot's model

(3) $C(\pi, \widehat{\mathcal{M}}_h^R) = C^*_{\widehat{\mathcal{M}}_h^R}$

→ *π is also optimal in the updated human model*



[IJCAI 2017; IJCAI 2019]



action: move p1 p2
p1
p2
p3
p4
p5
p6
p7
p8
p9
System Description
Plan Creation
Experiments
Arizona State University

Model Space Search for Model Reconciliation

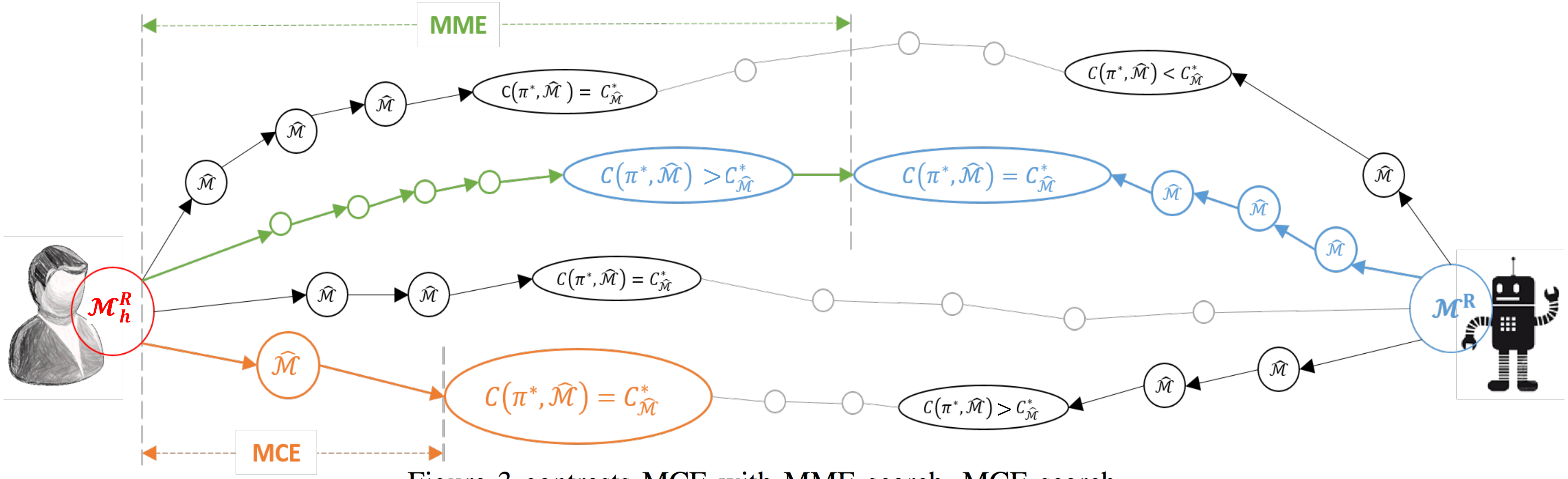
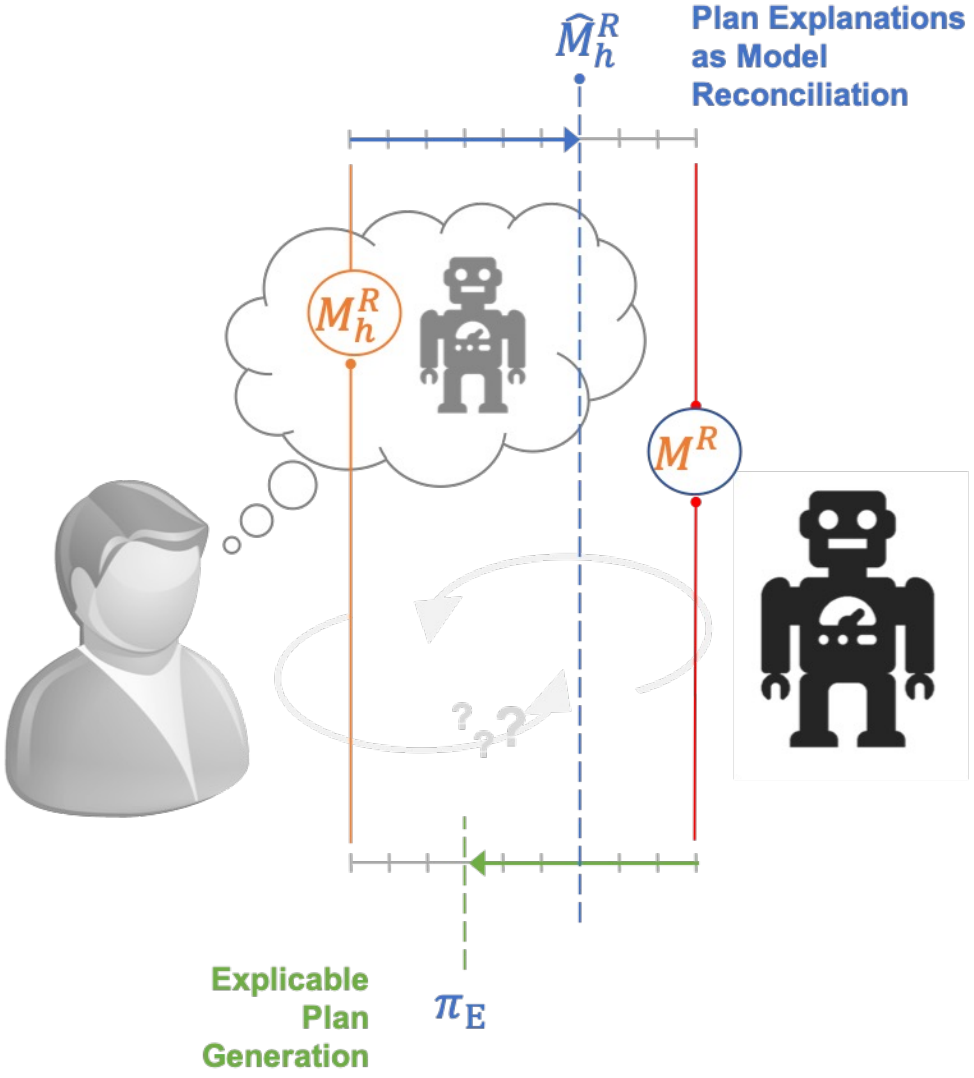


Figure 3 contrasts MCE with MME search. MCE search starts from $\mathcal{M}^H$, computes updates $\widehat{\mathcal{M}}$ towards $\mathcal{M}^R$ and returns the first node (indicated in orange) where $C(\pi^*, \widehat{\mathcal{M}}) = C^*_{\widehat{\mathcal{M}}}$. MME search starts from $\mathcal{M}^R$ and moves towards $\mathcal{M}^H$. It finds the longest path (indicated in blue) where $C(\pi^*, \widehat{\mathcal{M}}) = C^*_{\widehat{\mathcal{M}}}$ for all $\widehat{\mathcal{M}}$ in the path. The MME (shown in green) is the rest of the path towards $\mathcal{M}^H$.

Expectation-Aware Planning: A Unifying Framework

- Planning in the presence of external expectations
  - The Robot has both standard "**ontic**" actions and "**explanatory**" actions (e.g. speech acts)
  - It can model the effect of its actions on both its state, and the human's mental state
  - Planning is "multi-model"-–the AI agent uses both its model, and the human's expectation model to generate a course of action that contains both explanatory and ontic actions

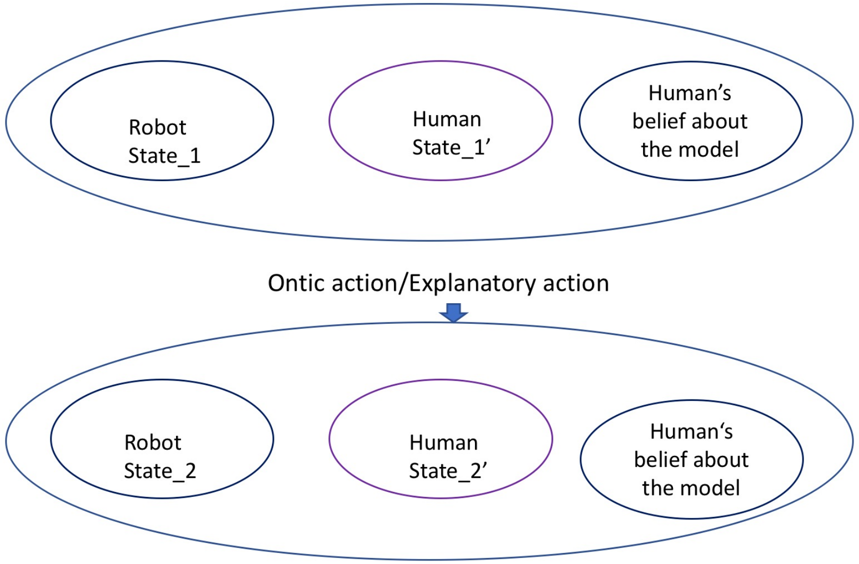


Expectation-Aware Planning

Expectation-aware planning problem consists $\Psi = \langle M_R, M_R^h \rangle$

$M^R = \langle F, A_R, I_R, G_R \rangle$ (Agent model)

$M_h^R = \langle F, A_R^h, I_R^h, G_R^h \rangle$ (User's model of the Agent)



The solution for a expectation-aware planning problem can be thought of as a tuple of the form $\langle E^\psi, \pi^\psi \rangle$

E^Ψ- (Explanatory) Information regarding the robot model

π^ψ- Agent plan

Such that

$$\pi^\psi \vDash_{M_h^R + E^\psi} G_H \text{ and } \pi^\psi \vDash_{M^R} G_R$$

Purely Explicable plan

$E^\psi = \phi$, π^ψ is executable in M_H

Explanation

π^ψ is selected for M_R and E^ψ information about M_R

Compilation to Classical (Single agent) Planning

(:action pickup

:parameters (?x - block)

:precondition (and (ontable ?x) (handempty))

:effect (and (not (ontable ?x)) (holding ?x))

)

Robot's model

(:action pickup

:parameters (?x - block)

:precondition (and (ontable ?x) ??)

:effect (and (not (ontable ?x)) (holding ?x))

)

Human's Expectation about the model

(:action pickup
:parameters (?x - block)
:precondition (and (ontable ?x) (handempty)
$\boxed{B_H\text{(ontable ?x)}}$
(implies (μ(pickup-has-precondition-handempty)) $\boxed{(B_H\text{(handempty))}}$)
:effect (and (not (ontable ?x) $\boxed{B_H\text{(ontable ?x))}}$ (holding ?x)
$\boxed{(B_H\text{(holding ?x))}}$)
)

Compilation

Meta-fluent controlled via an explanatory action

Human's belief about the state fluents

How does the AI Agent get the Human's Model?

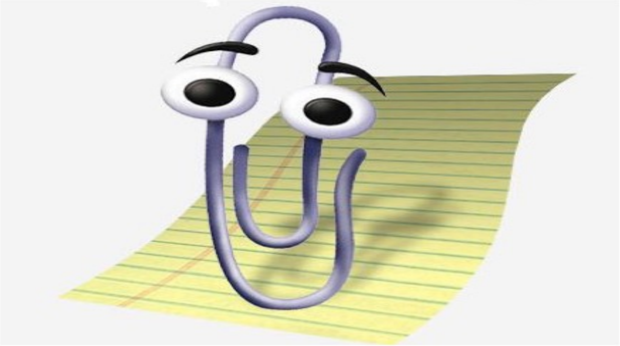
- In some cases (e.g. USAR scenario), the human and AI agent will start with the same shared model. All that is needed will be tracking the model drift
- Even if the robot doesn't know the model $\mathcal{M}_h^R$ with certainty, it can reason with multiple possible models [ICAPS 2018]
- In other cases, the AI agent does need to learn the human mental models [AAMAS 2015; AAMAS 2016]
  - Note however that while $\mathcal{M}_r^H$ can be learned from prior behavior traces of the human, $\mathcal{M}_h^R$ requires human's feedback on robot's behavior traces.
- Even when there are vocabulary differences between human and robot models, we can learn the human expectations rather than the actual model that results in those expectations
  - Model-free Explicability [ICRA 2017]
  - Model-free Explanation [IJCAI 2019]

M_r^H and $\widetilde{M}_h^R$ are *Expectations* on Models $\boldsymbol{\mathcal{M}}^H$ and $\boldsymbol{\mathcal{M}}^R$

They don't have to be executable

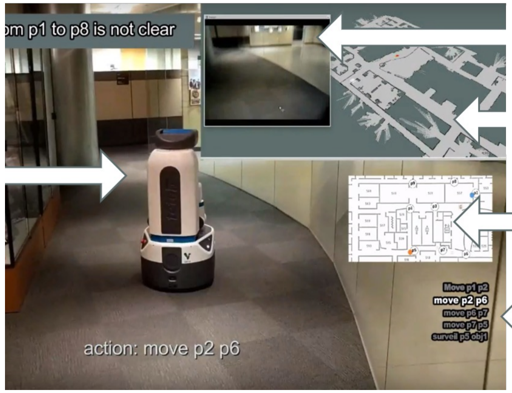
Do we really know what (sort of assistance) humans want?

Proactive Help Can be Disconcerting!



Solution: IRB-approved Systematic Human Subject Studies

Human-Factors Evaluation of the Model Reconciliation Process



Map used by the external

Plan currently under progress

Map available to the external

Plan of the internal agent

Explanations provided by the internal

Request for plan

Request for explanation

Evaluate plan

Map Annotations

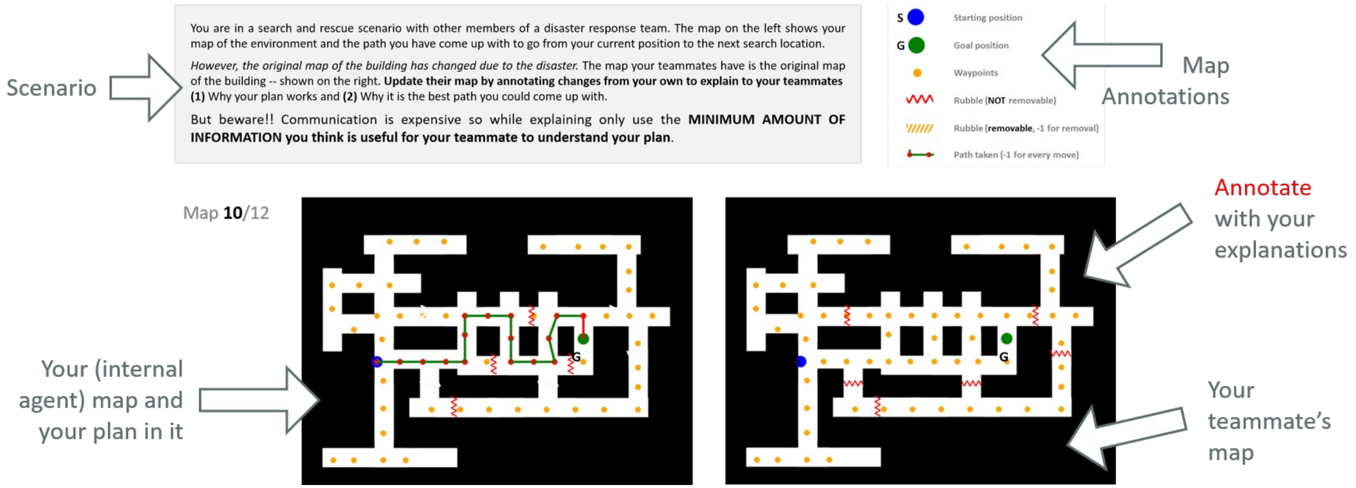
Score Board

Evaluate plan

[HRI 2019; HCI Journal 2020]

Human-Factors Evaluation of the Model Reconciliation Process

Case-1: How do humans explain the same scenarios?



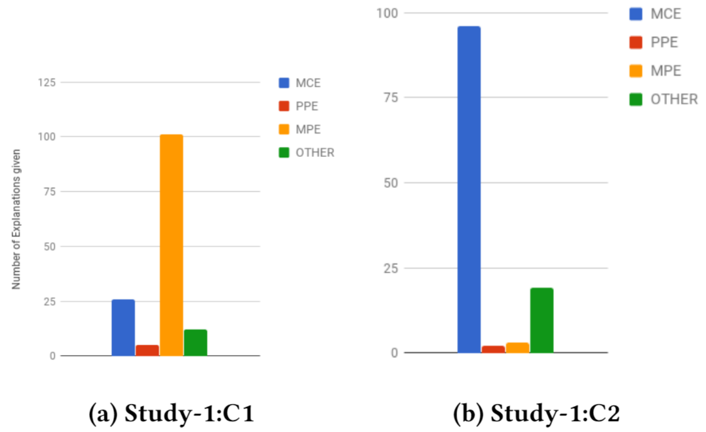


Figure 4: Different types of explanations for Study-1:C1;C2.

[HRI 2019; HCI Journal 2020]

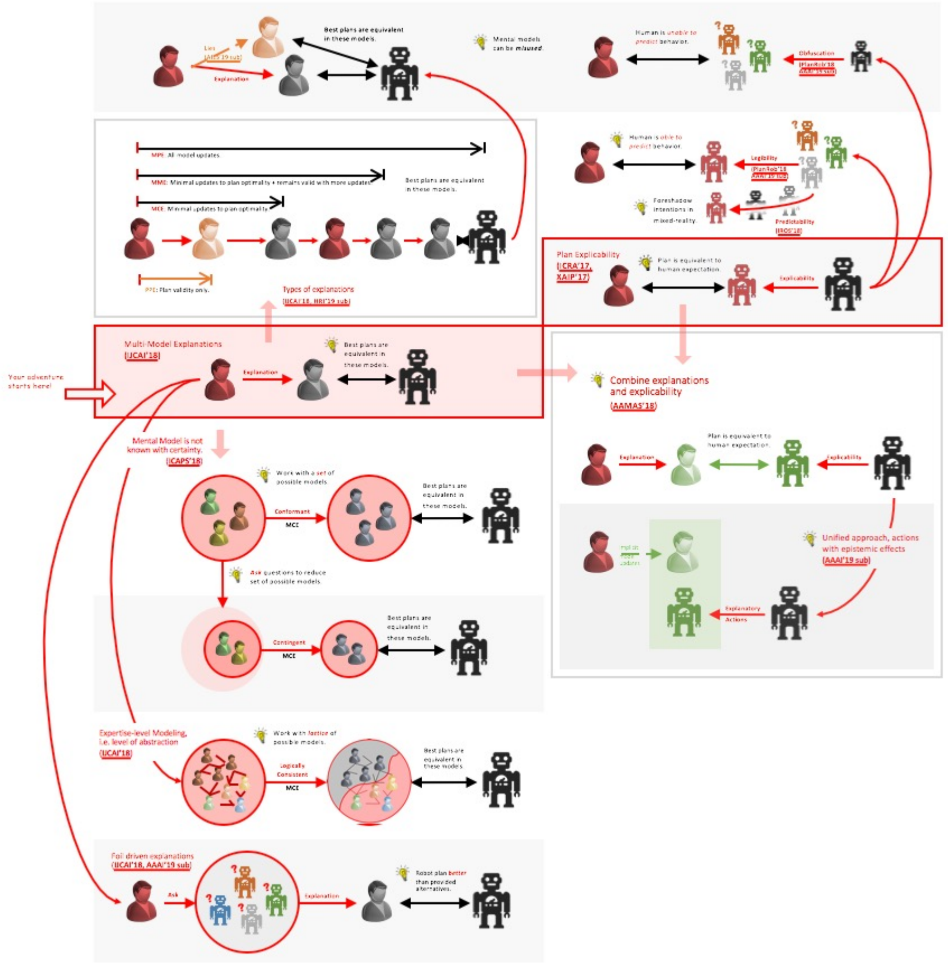
When (& Why) do Humans ask for Explanations from each other?

- When they are confused/surprised by the behavior (It is not what they *expected*--thus *inexplicable*).
  - Note that the confusion is orthogonal to "correctness"/"optimality" of the behavior. You may well be confused/surprised if your 2 year old nephew is able to give the exact distance between the Earth and the Sun.
    - $\mathcal{M}_h^R$ is too different from $\mathcal{M}^R$
  - Explanation here helps reconcile the expectations
    - Explanation is an attempt by the AI agent to get $\mathcal{M}_h^R$ closer to $\mathcal{M}^R$
- When they want to teach the other person and/or make sure that the decision was not a fluke and that the other person really understands the rationale for their decision.
  - Using the explanation to localize the fault, as it were..
- Note that the need for explanation is dependent on one person's model of the other person's capabilities/reasoning
  - Customized explanations (A doctor explains her decision to her patient in one way and to her doctor colleagues in a different way)
    - Explanation is needed when $\mathcal{M}_h^R$ (and not $\mathcal{M}^H$) is too different from $\mathcal{M}^R$; they are customized to $\mathcal{M}_h^R$
  - As the models get reconciled, there is less need for explanations in subsequent interactions!
- Explanations are connected to trust. We ask fewer explanations from people whom we trust

(There is also the whole "explanation of natural phenomena w.r.t scientific theories")

(Many) Extensions of the basic framework

- Supporting model reconciliation in non-PDDL settings [IJCAI 2019; ICAPS 2020]
- Relating other formulations of interpretable behavior [ICAPS 2019; IJCAI 2020]
- Handling foils & models at different levels of abstraction [IJCAI 2018]
    - Explaining unsolvability [IJCAI 2019]
- Handling multiple human agents [ICAPS 2018; IROS 2021]
    - Handling incomplete models; learning user types
- Implications to Trust & Deception
    - Mental modeling for obfuscation [AAAI 2019]
    - Lying with mental models [AIES 2019]
    - Engendering trust to improve performance [HRI 2023]



MORGAN & CLAYPOOL PUBLISHERS

Artificial Intelligence and Machine Learning >> Explainable Human-AI Interaction

Explainable Human-AI Interaction
A Planning Perspective

Sarath Sreedharan, Arizona State University,
Anagha Kulkarni, Arizona State University,
Subbarao Kambhampati, Arizona State University.
ISBN: 9781636392899 | **PDF ISBN:** 9781636392905

DOI: 10.2200/S01152ED1V01Y202111AIM050
Many institutions worldwide provide digital library access to Morgan & Claypool titles. You can check for personal access by clicking on the DOI link.

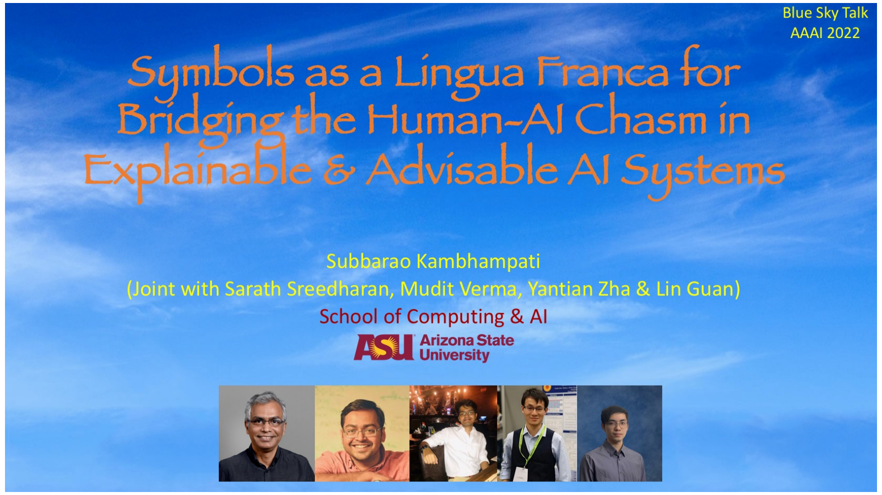
From its inception, artificial intelligence (AI) has had a rather ambivalent relationship with humans—swinging between their augmentation and replacement. Now, as AI technologies enter our everyday lives at an ever-increasing pace, there is a greater need for AI systems to work synergistically with humans. One critical requirement for such synergistic human-AI interaction is that the AI systems' behavior be explainable to the humans in the loop. To do this effectively, AI agents need to go beyond planning with their own models of the world, and take into account the mental model of the human in the loop. At a minimum, AI agents need approximations of the human's task and goal models, as well as the human's model of the AI agent's task and goal models. The former will guide the agent to anticipate and manage the needs, desires and attention of the humans in the loop, and the latter allow it to act in ways that are interpretable to humans (by conforming to their mental models of it), and be ready to provide customized explanations when needed.

The authors draw from several years of research in their lab to discuss how an AI agent can use these mental models to either conform to human expectations or change those expectations through explanatory communication. While the focus of the book is on cooperative scenarios, it also covers how the same mental models can be used for obfuscation and deception. The book also describes several real-world application systems for collaborative decision-making that are based on the framework and techniques developed here. Although primarily driven by the authors' own research in these areas, every chapter will provide ample connections to relevant research from the wider literature. The technical topics covered in the book are self-contained and are accessible to readers with a basic background in AI.

https://bit.ly/3GeU2Dx

Talk Overview

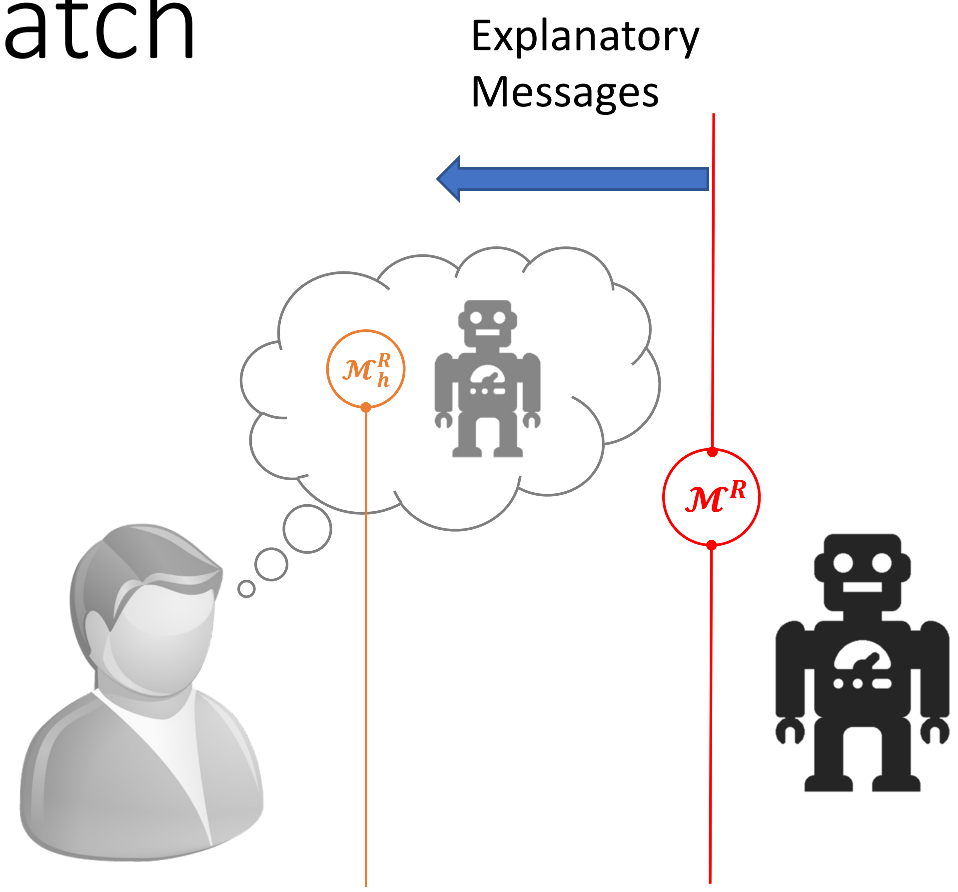
- Part 1: Why and how do humans exchange explanations? Do AI systems need to?
- Part 2: Using Mental Models for Explainable Behavior in the context of explicit knowledge tasks (think Task Planning)
  - The 3-model framework: $\mathcal{M}^R, \mathcal{M}^H, \mathcal{M}_h^R$
  - Explicability: *Conform to* $\mathcal{M}_h^R$
  - Explanation: Reconcile $\mathcal{M}_h^R$ to $\mathcal{M}^R$
  - Extensions: *Foils, Abstractions, Multiple Humans..*
- Part 3: Supporting explainable behavior even without shared vocabulary
  - Symbols as a *Lingua Franca* for Explainable and Advisable Human-AI Interaction
    - *Post hoc symbolic explanations of inscrutable reasoning*
    - *Accommodating symbolic advice into inscrutable systems*



Addressing Vocabulary Mismatch

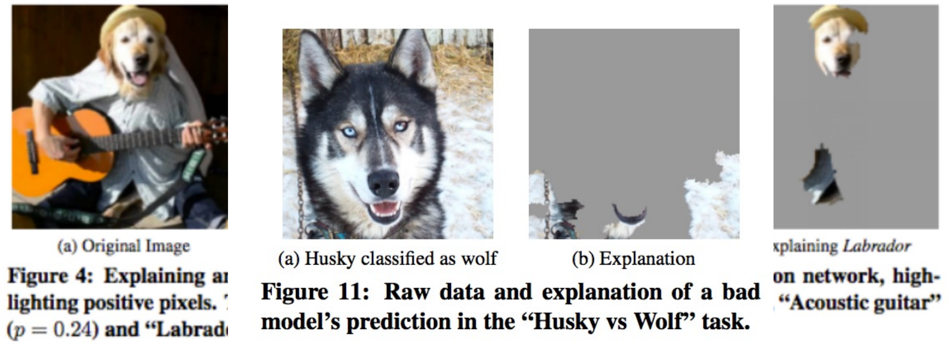
- We assumed a shared vocabulary as a starting point

Agent may be using a learned model or an inscrutable simulator



Explanations in the absence of shared vocabulary

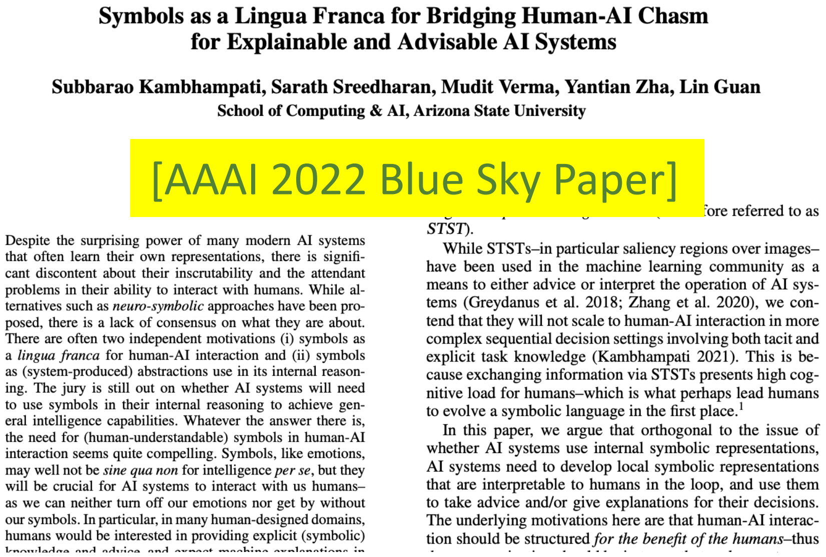
- What about explanations in the absence of shared vocabulary?
  - E.g. AI agents working off of their own internal learned representations?
- The lowest common denominator between humans and the AI agents in such cases will be just raw signals and data
  - Explanations in terms of them will involve exchanging (or "pointing to") "Space Time Signal Tubes" (STSTs)
  - Interestingly, this is what a majority of XAI literature does!
- "XAI" is hot.. But mostly as a [debugging tool]() for "inscrutable" representations
  - "Pointing" explanations (primitive)
    - Explaining decisions will involve pointing over space-time signal tubes!



Please point to the "ostrich" parts

``Pointing Explanations” are hard to comprehend!

- Pointing explanations with STSTs are not only unwieldy (in terms of communication costs), but also **hard to comprehend in many cases**
- Humans (1) develop a shared symbolic vocabulary and (2) exchange symbolic explanations where possible, and (3) come down to pointing explanations *only when the vocabulary is inadequate* (and use this as a sign to *expand vocabulary)*
  - This approach works particularly well for explicit knowledge tasks (but we also use it for mixed and tacit-knowledge tasks—think of “pick and roll” in basketball)
- We advocate a symbolic interface layer instead..



Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems

Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, Lin Guan
School of Computing & AI, Arizona State University

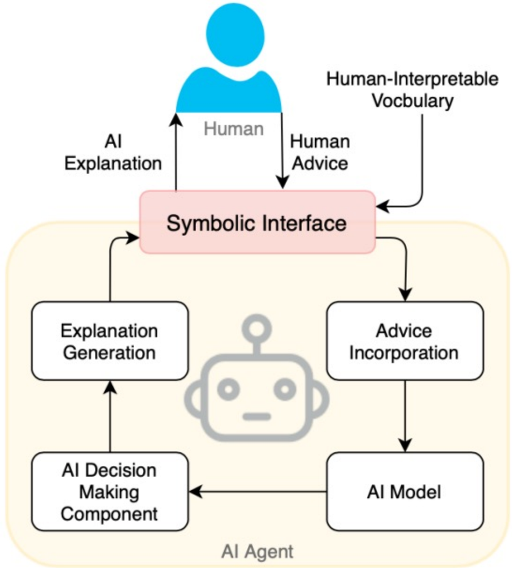
[AAAI 2022 Blue Sky Paper]

Despite the surprising power of many modern AI systems that often learn their own representations, there is significant discontent about their inscrutability and the attendant problems in their ability to interact with humans. While alternatives such as *neuro-symbolic* approaches have been proposed, there is a lack of consensus on what they are about. There are often two independent motivations (i) symbols as a *lingua franca* for human-AI interaction and (ii) symbols as (system-produced) abstractions use in its internal reasoning. The jury is still out on whether AI systems will need to use symbols in their internal reasoning to achieve general intelligence capabilities. Whatever the answer there is, the need for (human-understandable) symbols in human-AI interaction seems quite compelling. Symbols, like emotions, may well not be *sine qua non* for intelligence *per se*, but they will be crucial for AI systems to interact with us humans–as we can neither turn off our emotions nor get by without our symbols. In particular, in many human-designed domains, humans would be interested in providing explicit (symbolic)

...fore referred to as *STST*).

While STSTs–in particular saliency regions over images–have been used in the machine learning community as a means to either advice or interpret the operation of AI systems (Greydanus et al. 2018; Zhang et al. 2020), we contend that they will not scale to human-AI interaction in more complex sequential decision settings involving both tacit and explicit task knowledge (Kambhampati 2021). This is because exchanging information via STSTs presents high cognitive load for humans–which is what perhaps lead humans to evolve a symbolic language in the first place.[1]

In this paper, we argue that orthogonal to the issue of whether AI systems use internal symbolic representations, AI systems need to develop local symbolic representations that are interpretable to humans in the loop, and use them to take advice and/or give explanations for their decisions. The underlying motivations here are that human-AI interaction should be structured *for the benefit of the humans*–thus



AI systems must be Explainable and Advisable

- As we are increasingly surrounded by AI systems, it is critical that they are explainable and advisable
- The explainability and advisability must be on *our (human)* terms
  - We shouldn't have to debug AI systems to interpret them
    - It would be a pity if all the progress in AI results in us humans going into the (incomprehensible) land of the AI systems
  - We want them to communicate with us in our terms
- We argue that AI systems need to support a *symbolic lingua franca* with the humans in the loop

Neuro-Symbolic AI: Two orthogonal motivations

Internal Symbolic reasoning

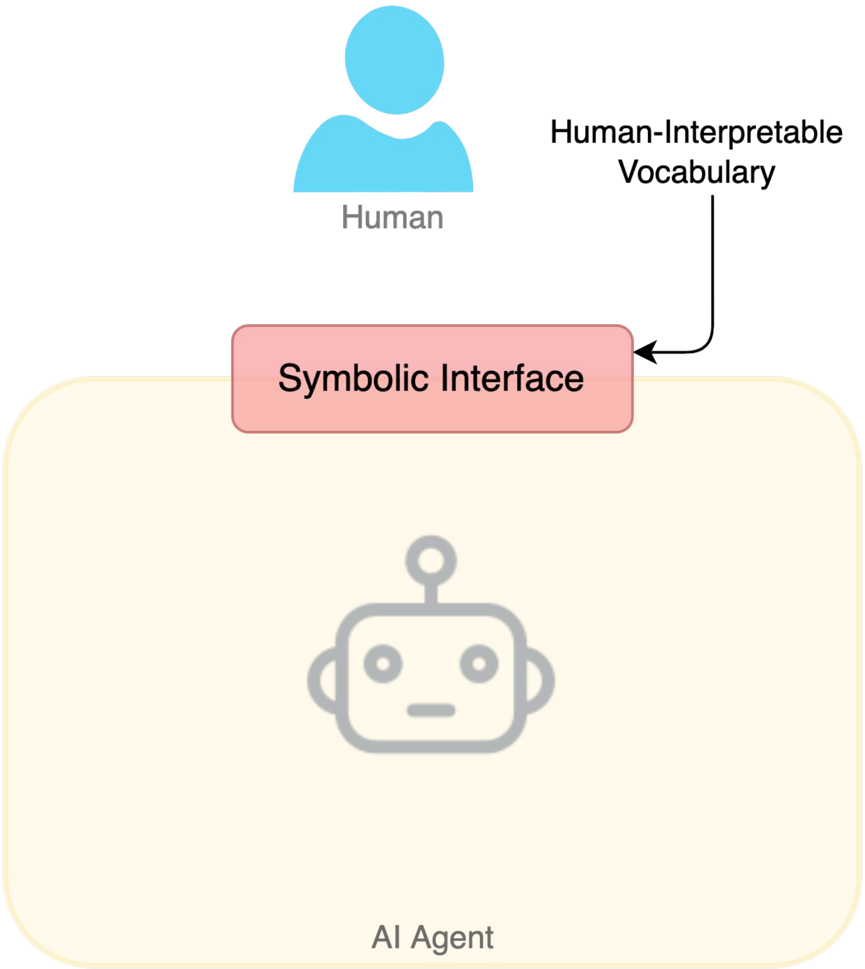
- Argument that AI systems would need to employ internal symbolic reasoning for efficiency & scalability
  - The jury is still very much out on this
- (There is little reason to expect that symbols used as abstractions in internal reasoning will align well with those that humans use)

Symbolic communication interface

- Argument that (regardless of their internal reasoning modality), AI systems must support a symbolic communication channel with humans (using symbols that make sense to humans)
    - The alterative—of exchanging Space Time Signal Tubes (STSTs)—presents intolerably high cognitive load for humans!
  - This *Symbolic Lingua-Franca for explainability and advisability is* the main argument of our paper
    - This may well be *in addition* to other modalities of communication

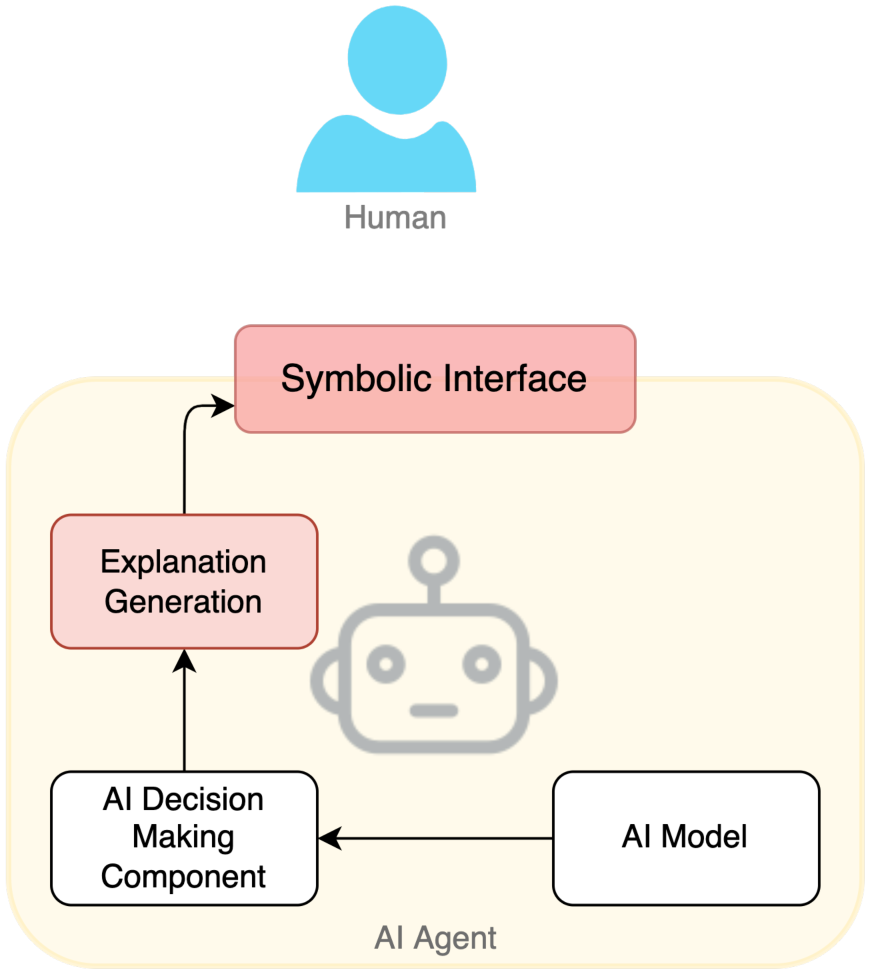
Use case for the Symbolic Layer

- We will be using the shared vocabulary to build an approximate symbolic representation of agent model that is surfaced to the user
- The symbolic model aims to capture the human's understanding of the robot model -- M_h^R
  - It can thus be used as the basis for any human-robot interaction that depends on M_h^R
- In particular, we can use this symbolic interface for
  - Generating Explanations
  - Accept advice from the user



Generating Explanation

- We can use the symbolic model as the basis for explaining any decisions made by the system
- We can directly leverage this model in the context of the model-reconciliation framework developed for symbolic models.
- The symbolic model, being an approximation of the underlying system model, may be insufficient to explain all the system decisions – as such explanation may require expanding the symbolic model to provide sufficient explanation
    - A special case of model-reconciliation where there is an additional translation process



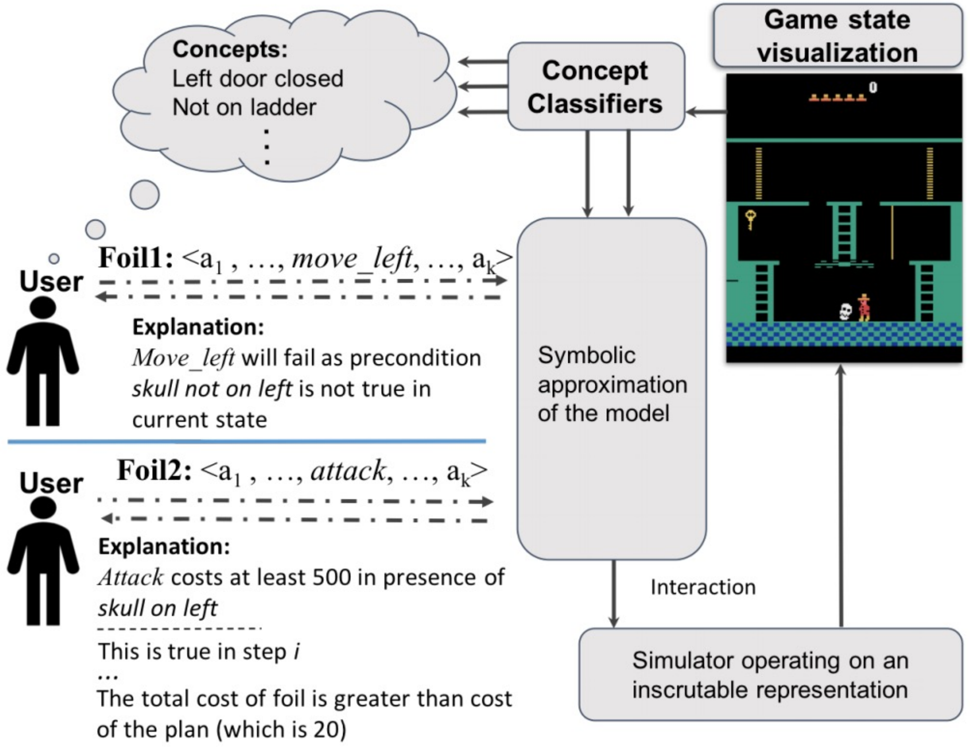
Explaining In terms of User Specified Concepts

User specifies concepts

-- Each concept maps to a binary classifier

User raises a foil – i.e., an alternate plan – A model component learned to refute the foil

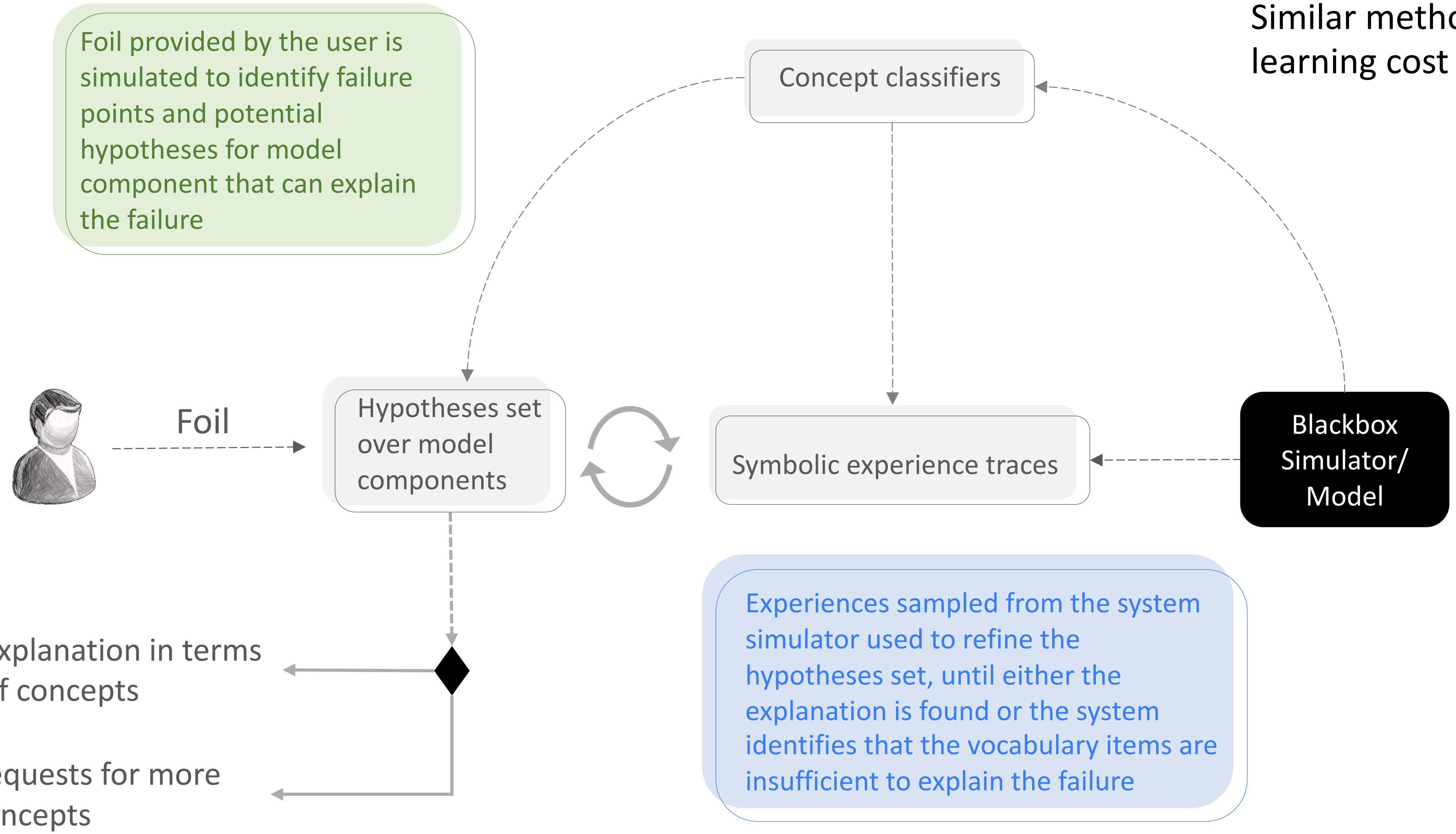
The foil fails at any point	*Identify the missing preconditions*
The foil is costlier than the original plan	*Identify an abstract version of the cost function*



[ICLR, 2022]

Learning Model Components Through Sampling

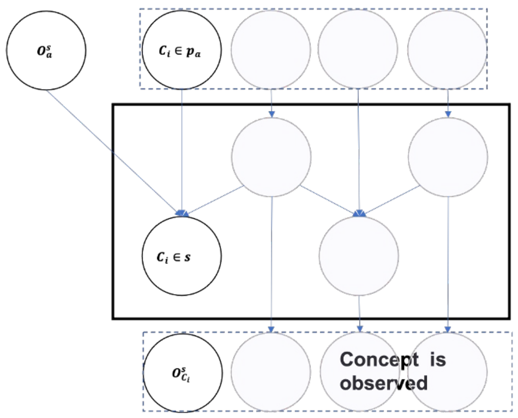
Similar method can be used for learning cost function



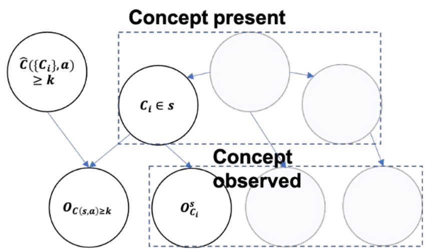
Generating Confidence-level For the Explanation

- Generate confidence to account for
  - Sampling based generation method
  - Noisiness of classifiers used to generate explanations

- Avoid creating explanations that build undeserved trust in the system



Graphical model for Calculating Posterior Probability of a concept being a precondition



Graphical model for Calculating Posterior Probability of a concept being part of the cost function

Empirical Evaluation

Table 1: Results from the user study

	Prefers symbols	Average Likert-score	P-value
Precondition	19/20	3.47	1.0×10^{-8}
Cost	16/20	3.21	0.03

(a) H1

Method	# of Participant	Average Time Taken (sec)	Average # of Steps
Concept-Based	23	43.78 ± 12.59	35.87 ± 9.69
Saliency Map	25	134.24 ± 61.72	52.64 ± 11.11

(b) H2

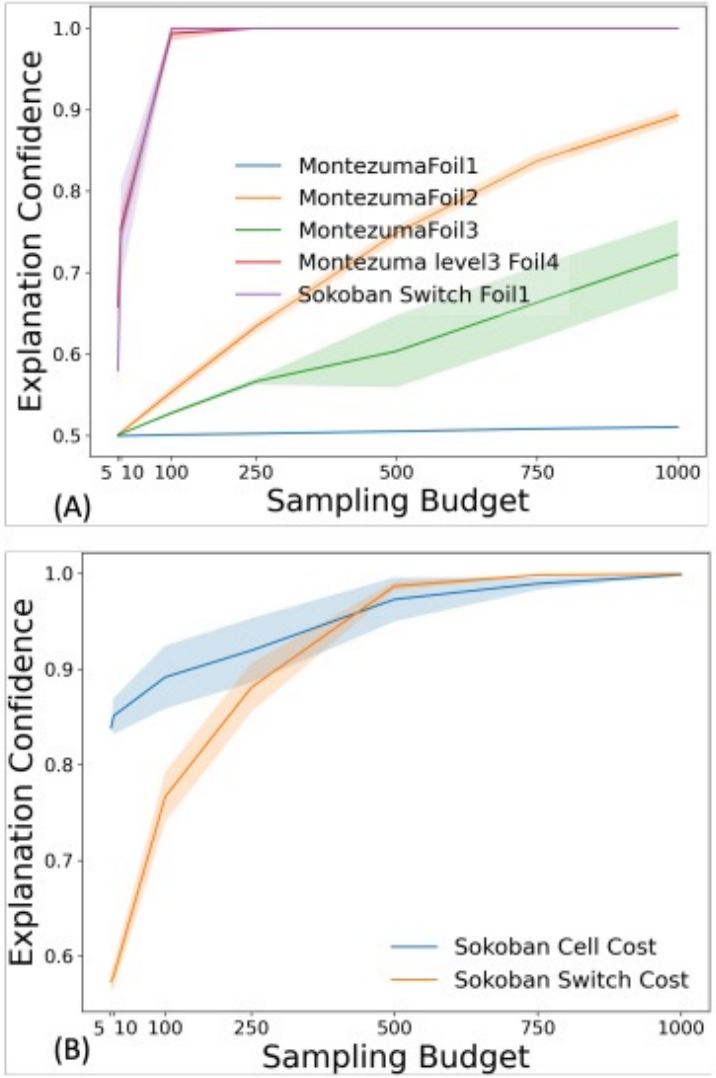
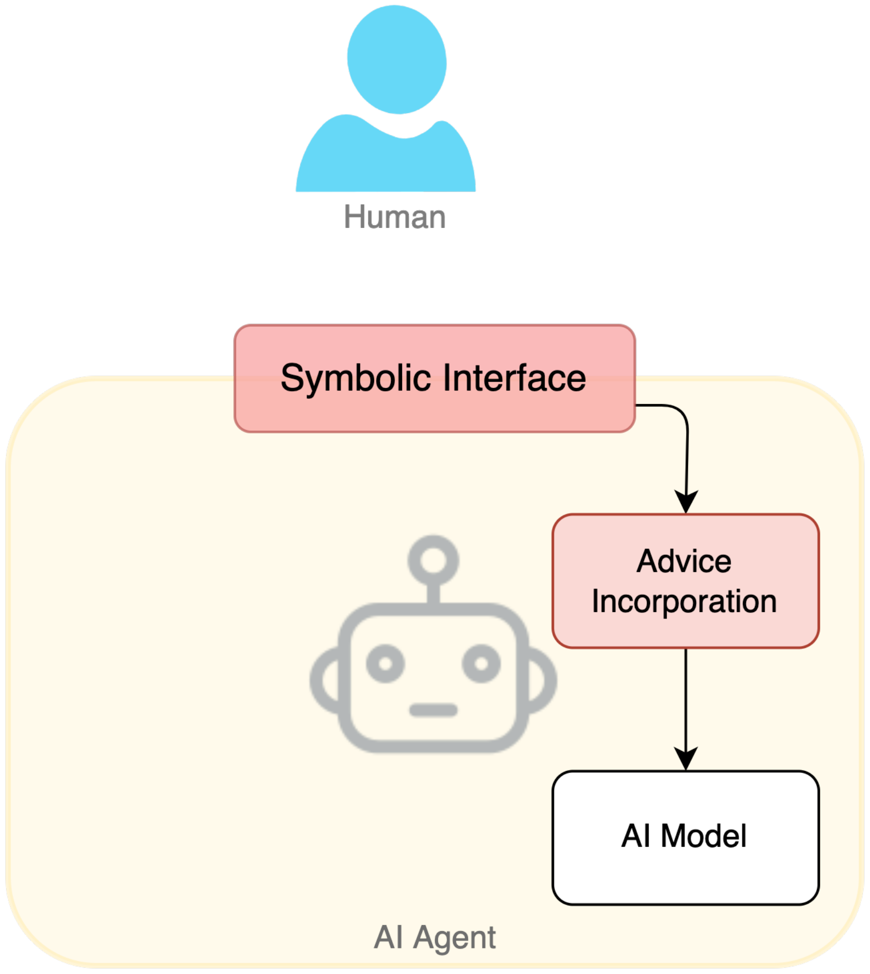


Figure 5: The average probability assigned to the correct model component by the search algorithms, calculated over ten random search episodes with std-deviation.

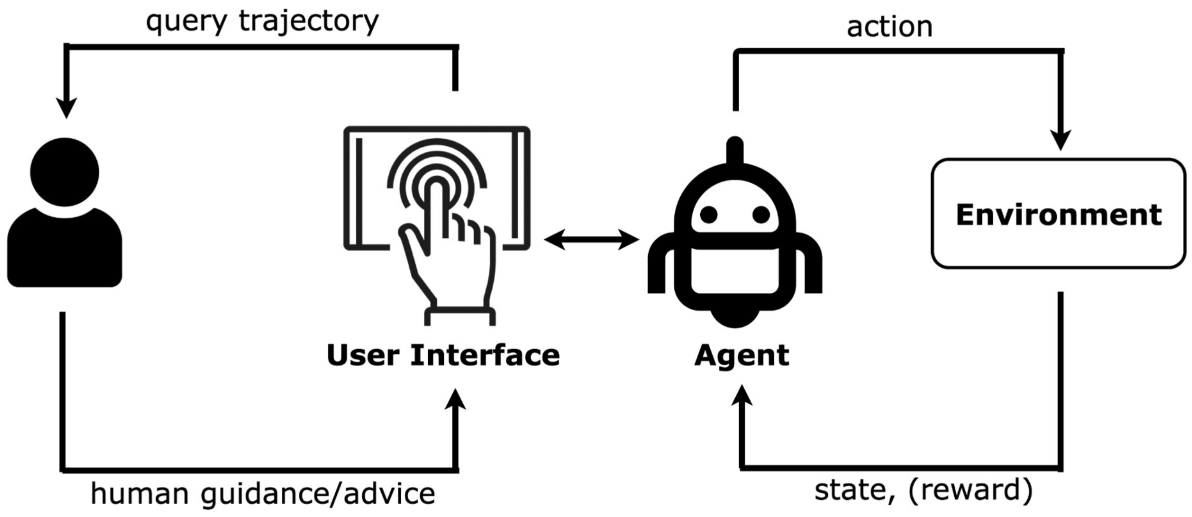
Accepting Advice

- The human user can directly update the model to drive system behavior
- The modifications made or constraints applied in the symbolic model are translated into a form that can be used by the low-level agent
  - The advise can either be given during the learning time (where the RL agent specifically requests for criticism)
    - [NeurIPS 2021 Spotlight]
  - Or before the RL phase starts—via a possibly incomplete symbolic model
    - [ICML 2022]
- Additionally, we can use the symbolic model as a basis to interpret even non-symbolic advice (e.g. demonstrations) provided by the user
  - For example, one could use the symbols and the model definition to better interpret input like human demonstration.



Human-Advisable RL

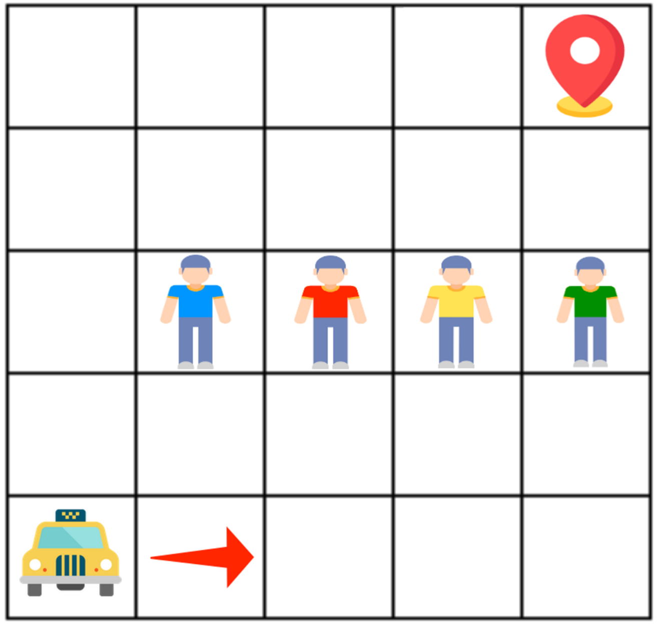
- A human trainer monitors the learning process of RL
- The agent adjusts its policy according to human advice
- Forms of advice
  - Inexpensive and intuitive to specify.
  - Reduced to TAMER [Knox and Stone, 2009] when advice is binary evaluative feedback
- Human-Advisable RL generalizes from Human-in-the-Loop RL (HIRL) but has separate challenges **beyond** HIRL

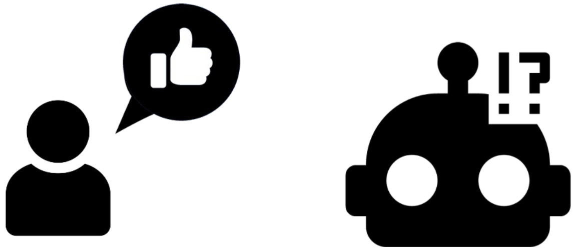


[NeurIPS 2021 Spotlight]

Challenges in Human-Advisable RL

- The Quandary:
  - Human feedbacks are **expensive** and **sparse**
  - DNNs are always **data-hungry**
- Missing **Lingua Franca** (shared vocabulary) between humans and agents
  - Limit the forms of feedback to **simple numerical labels** (e.g. evaluative feedback, binary preference labels)
  - Numerical labels are **not informative** enough
- Communicative Modalities
  - Humans prefer multi-modal communications
  - Easy (**effortless**) to provide
  - The agent can easily understand

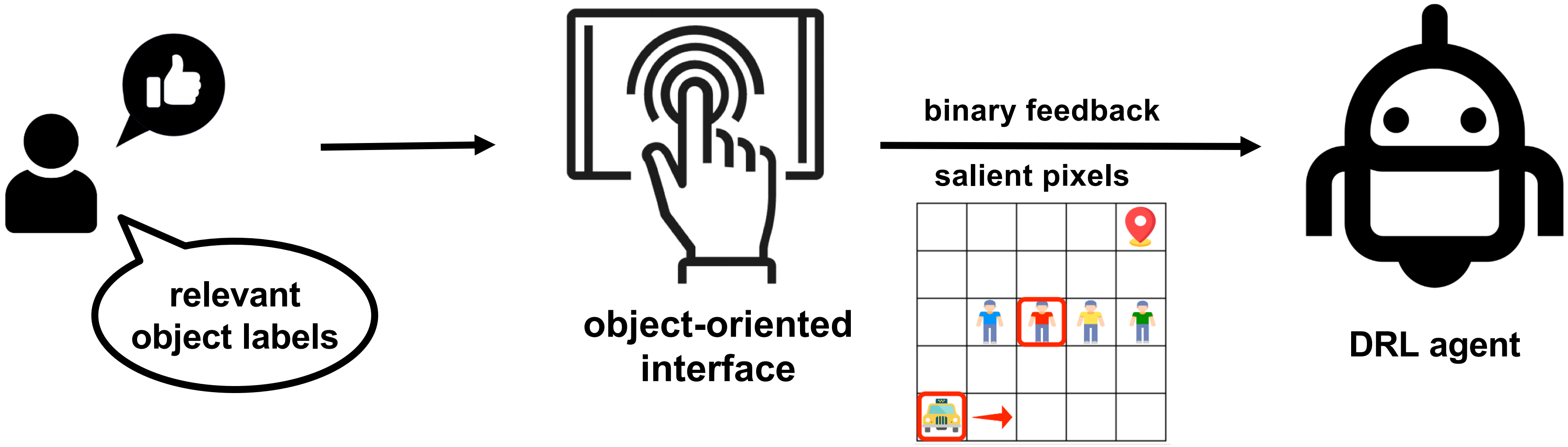




Binary feedback doesn't indicate why certain action is good/bad.

Our Goals

- The Quandary:
  - Improve **human feedback sample efficiency** & **environment sample efficiency**
- Lingua Franca & Multi-Modal Communication
  - Augment binary evaluative feedback with **human visual explanation**
    - Annotations of **task-relevant regions (pixels)** in image
    - Help in "maximally" utilizing each binary feedback
  - **Effortlessly** collect human visual feedback
    - An **object-oriented** middle layer (interface)



Efficiently Collecting Visual Explanation

An **object-oriented interface**:

- Observations:
  - Human visual explanations are usually associated with certain **objects** or **regions** in image
  - Salient regions/objects are usually the same in nearby frames
- Use a simple **tracking and detection** module to detect possible salient objects/regions
- **Effortless** communication at the level of **symbols** (e.g. object labels) even though the DRL agent is operating in pixel-space
- User study: collected over 2k feedbacks (binary feedback & visual explanation) in 30 min

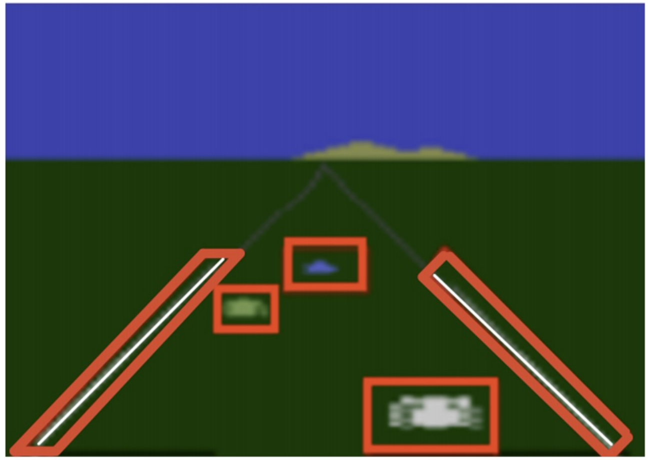
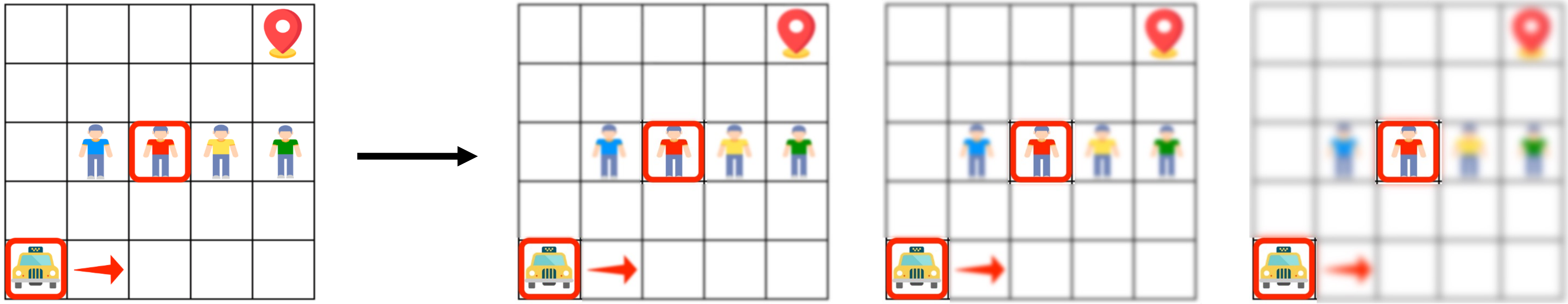


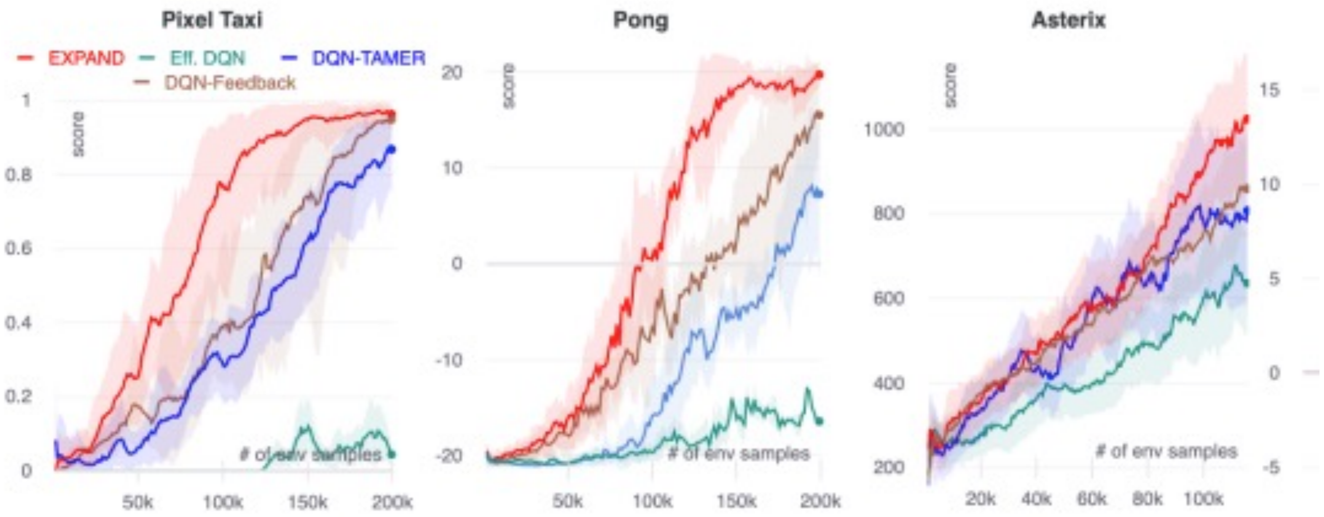
Fig. 3. All the lanes and cars are automatically highlighted and tracked, so the human trainers only need to deselect irrelevant objects in the image.

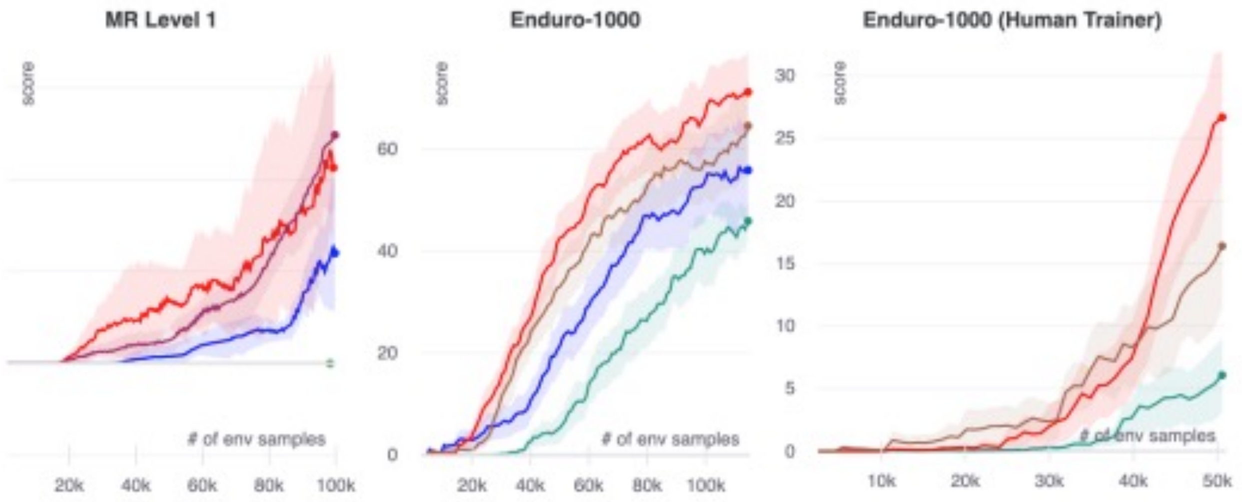
Context-Aware Data Augmentation

- Existing ways to incorporate saliency information into supervised learning systems are not suitable for **less stable** learning systems like deep reinforcement learning
- **Context-Aware Data Augmentation**
  - Intuition: small perturbations on irrelevant regions should not alter the agent's policy
  - Approach:
    - Apply various image transformations to the irrelevant regions, and obtain a set of augmented feedback
    - **Gaussian blurring** with different Gaussian kernels
    - Two loss terms to enforce invariance
  - Examples:

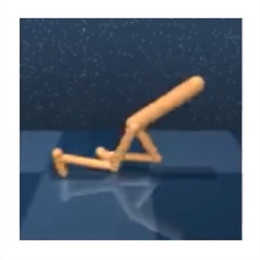


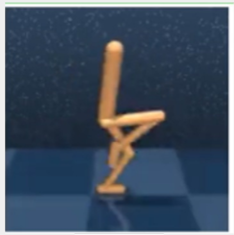
Experimental Results





Reward Learning from Trajectory Comparisons





Learn to give higher rewards to trajectories preferred by the human:

$$\sum_{s_t, a_t \in \sigma_0} r_\theta(s_t, a_t) < \sum_{s_t, a_t \in \sigma_1} r_\theta(s_t, a_t)$$

It assumes the objective can't be expressed in terms of nameable concepts.

Most suitable for **tacit-knowledge** tasks like learning locomotions

But need hundreds of preference labels!

Tweaking Agent Behavior through Relative Behavioral Attributes

- Allow users to specify the behavior through **explicit symbolic** concepts.
- Uses a parametric method to learn the **tacit** parts (e.g., how to walk naturally)

Our method

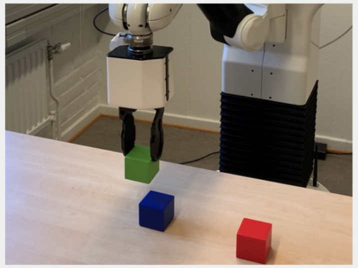
- Not good enough
- Move **more softly**/sneakily!

Updated behavior

Only need a small number of attribute feedback!

A natural way for human-agent communication

Symbolic Goal Specification



Example symbolic reward function:

$$r(s,a) = \begin{cases} 1 & \text{if Green is on Blue} \\ 0 & \text{otherwise} \end{cases}$$

Very straightforward and intuitive to use

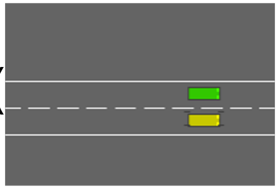
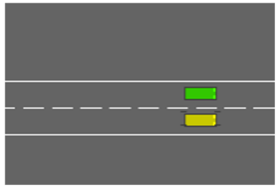
But limited to **explicit-knowledge** tasks (e.g., it's unclear how to define the ways of walking "charmingly" or "sneakily")

Relative Behavioral Attributes: An Example Method

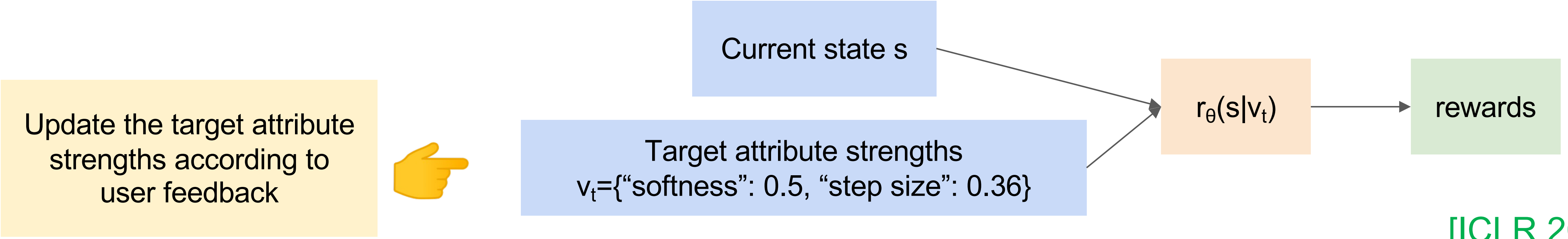
- Given a large-scale offline behavior datasets (e.g., Waymo driving dataset or human motion dataset), learn an **attribute-conditioned ranking function** (labels given by agent builder)

$f_\sigma($  $|$ "walk softly"$)$ $\prec$ $f_\sigma($  $|$ "walk softly"$)$

$f_\sigma($  $|$ "step size"$)$ $\prec$ $f_\sigma($  $|$ "step size"$)$

$f_\sigma($  $|$ "steering sharpness"$)$ $\prec$ $f_\sigma($  $|$ "steering sharpness"$)$

- Learn an **attribute parameterized reward function** (i.e., essentially a family of rewards that correspond to behaviors with diverse attribute strengths)



[ICLR 2023]

q=1

Attribute strength
Step size: 0.67
Softness: 0.51

q=2

Attribute strength
Step size: 0.70
Softness: 0.26

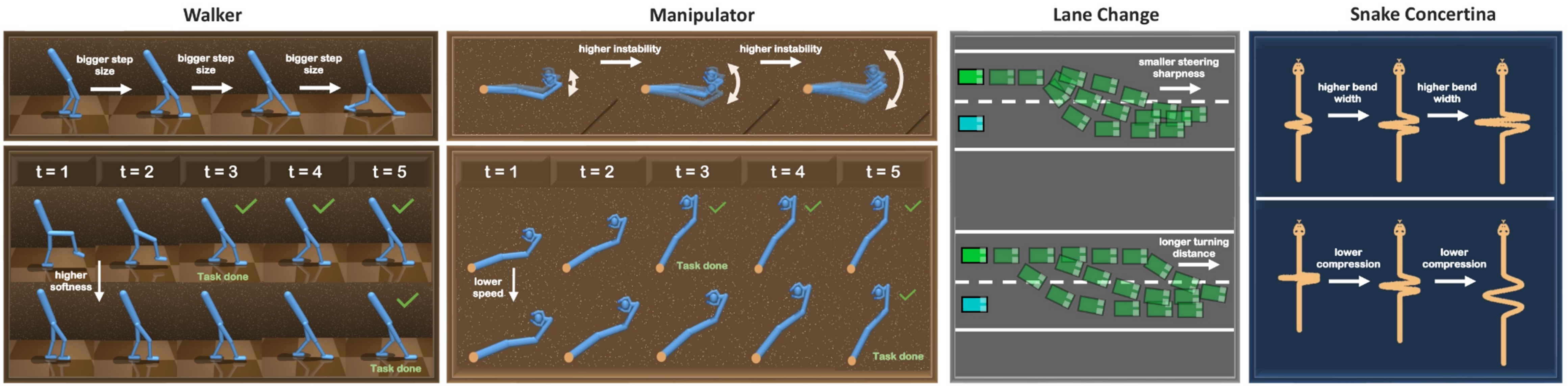
q=3

Attribute strength
Step size: 0.86
Softness: 0.16

q=4

Attribute strength
Step size: 0.79
Softness: 0.22

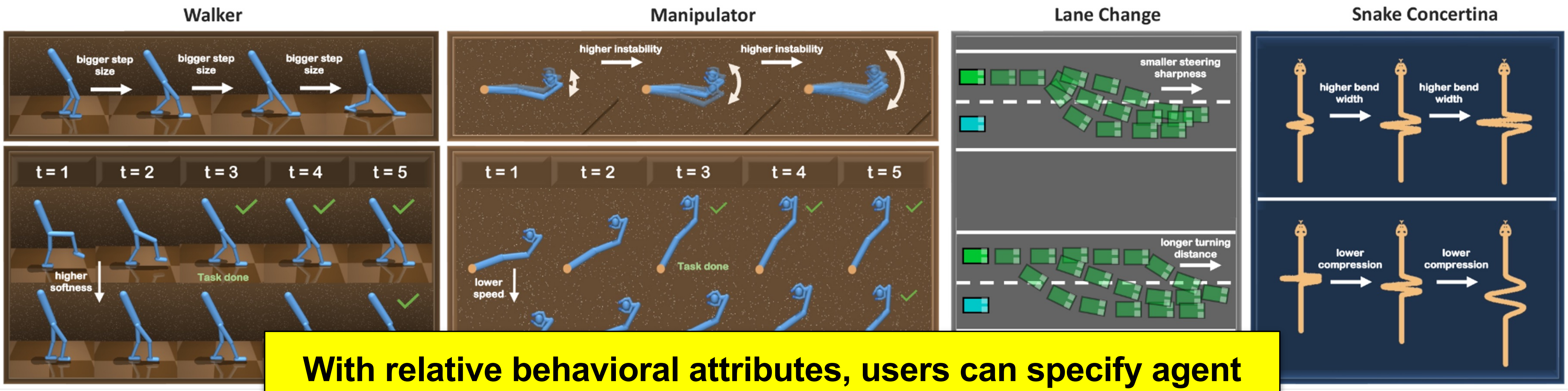
Results



Method	Lane-Change		Manipulator		Snake		Walker	
	SR	AF (std)	SR	AF (std)	SR	AF (std)	SR	AF (std)
RA-Global	0.95	3.95 *(2.43)*	1.0	2.8 *(1.21)*	**0.85**	**4.17 *(1.85)***	**1.0**	**3.75 *(1.47)***
RA-Global-L	**1.0**	**3.05 *(2.06)***	**1.0**	**2.5 *(1.32)***	0.8	6.38 *(5.03)*	0.95	3.78 *(2.25)*
PbRL	1.0	162.3 *(184)*	0.6	159.5 *(188.87)*	0.05	N/A	1.0	84.6 *(79.87)*

SR - Success Rate; AF - Average Feedback (when success); RA - Relative Attribute; L - Language

Results



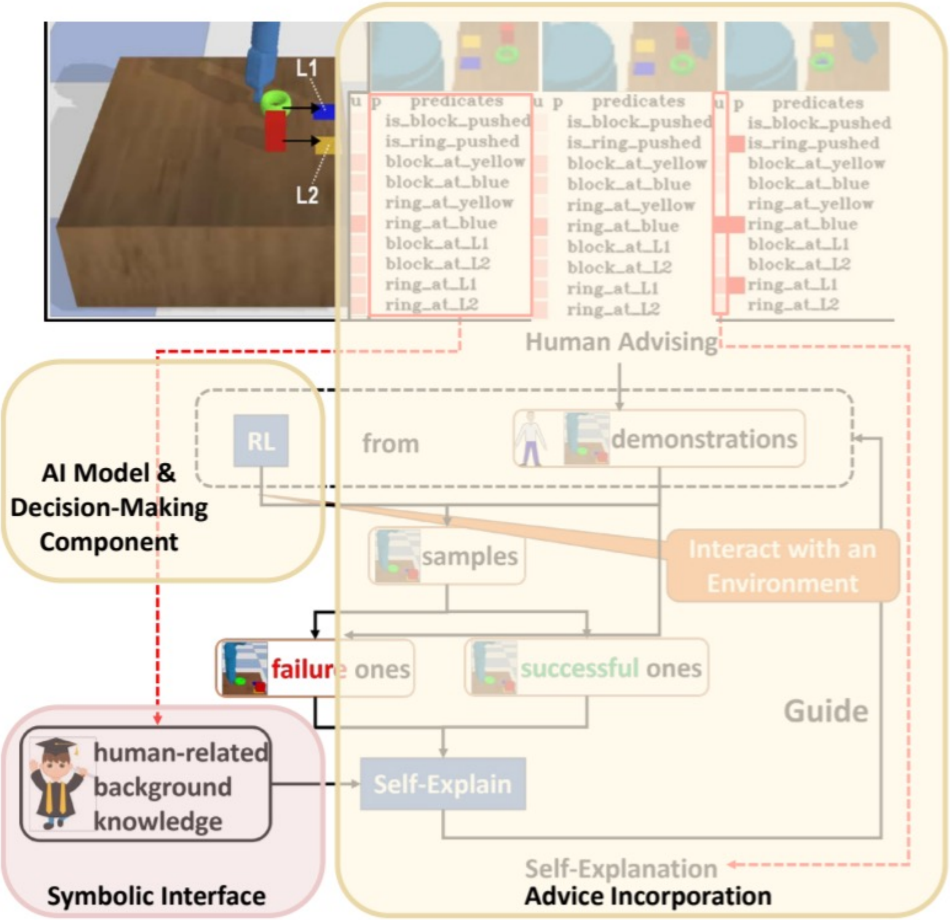
With relative behavioral attributes, users can specify agent behavior extremely efficiently!!

Method	Lane-Change		Manipulator		Snake		Walker	
	SR	AF (std)	SR	AF (std)	SR	AF (std)	SR	AF (std)
RA-Global	0.95	3.95 *(2.43)*	1.0	2.8 *(1.21)*	**0.85**	**4.17 *(1.85)***	**1.0**	**3.75 *(1.47)***
RA-Global-L	**1.0**	**3.05 *(2.06)***	**1.0**	**2.5 *(1.32)***	0.8	6.38 *(5.03)*	0.95	3.78 *(2.25)*
PbRL	1.0	162.3 *(184)*	0.6	159.5 *(188.87)*	0.05	N/A	1.0	84.6 *(79.87)*

SR - Success Rate; AF - Average Feedback (when success); RA - Relative Attribute; L - Language

Interpreting Ambiguous Human Demonstrations in terms of shared symbols

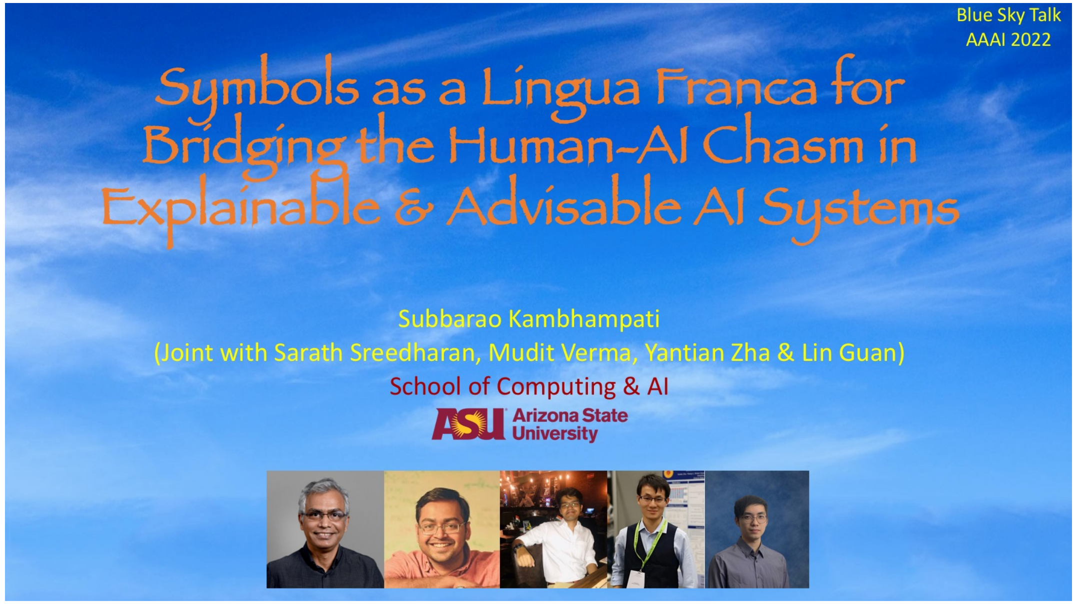
- SERLfD system leverages the symbolic interface to better interpret ambiguous human demonstrations
- System assumes that the (continuous) demonstration provided by the human is guided by their own interest in highlighting specific symbolic goals and way points.
    - It learns to interpret the relative importance of these symbols and use that to disambiguate the demonstrations
        - (Can be viewed as an exercise by the AI system to parse/explain the demonstration in terms of the shared symbols)



[AAAI 2022 Wkshp on RL in Games]

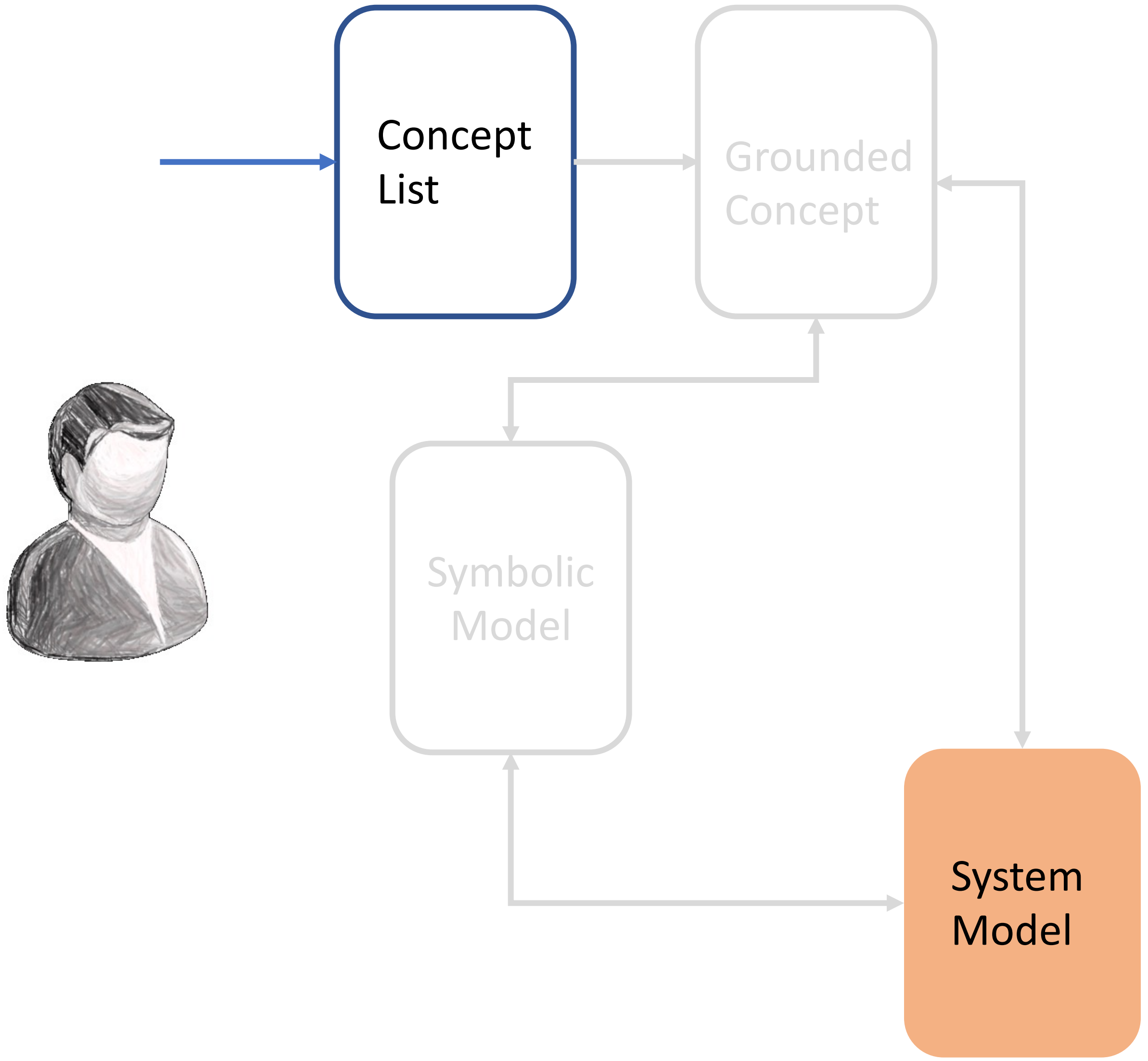
Open Research Challenges in Supporting Symbolic Interfaces

- Collecting initial concept set
- Grounding concept set
- Vocabulary expansion



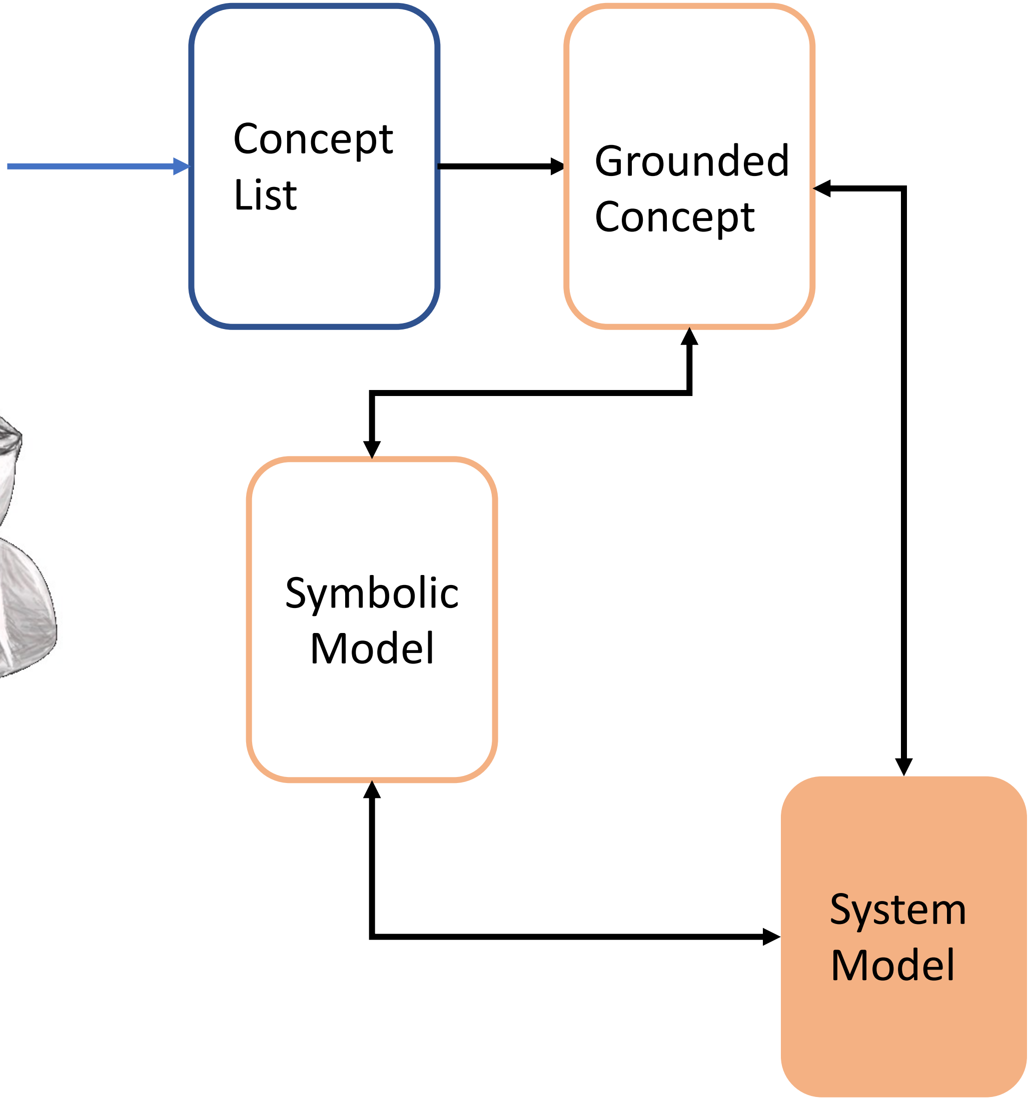
Challenge 1: Collecting Initial Concept Set

- Collect a set of propositional/relational concepts that will be used to build the symbolic interface
  - Captures a set of concepts that the human associates with the task
  - Each slice of STST meant to map to a set of these concepts
- For common tasks, one could leverage systems like **scene graph** analysis
  - The cost of concept acquisition amortized across multiple tasks
- Concepts could also be potentially mined from domain-specific databases/documents



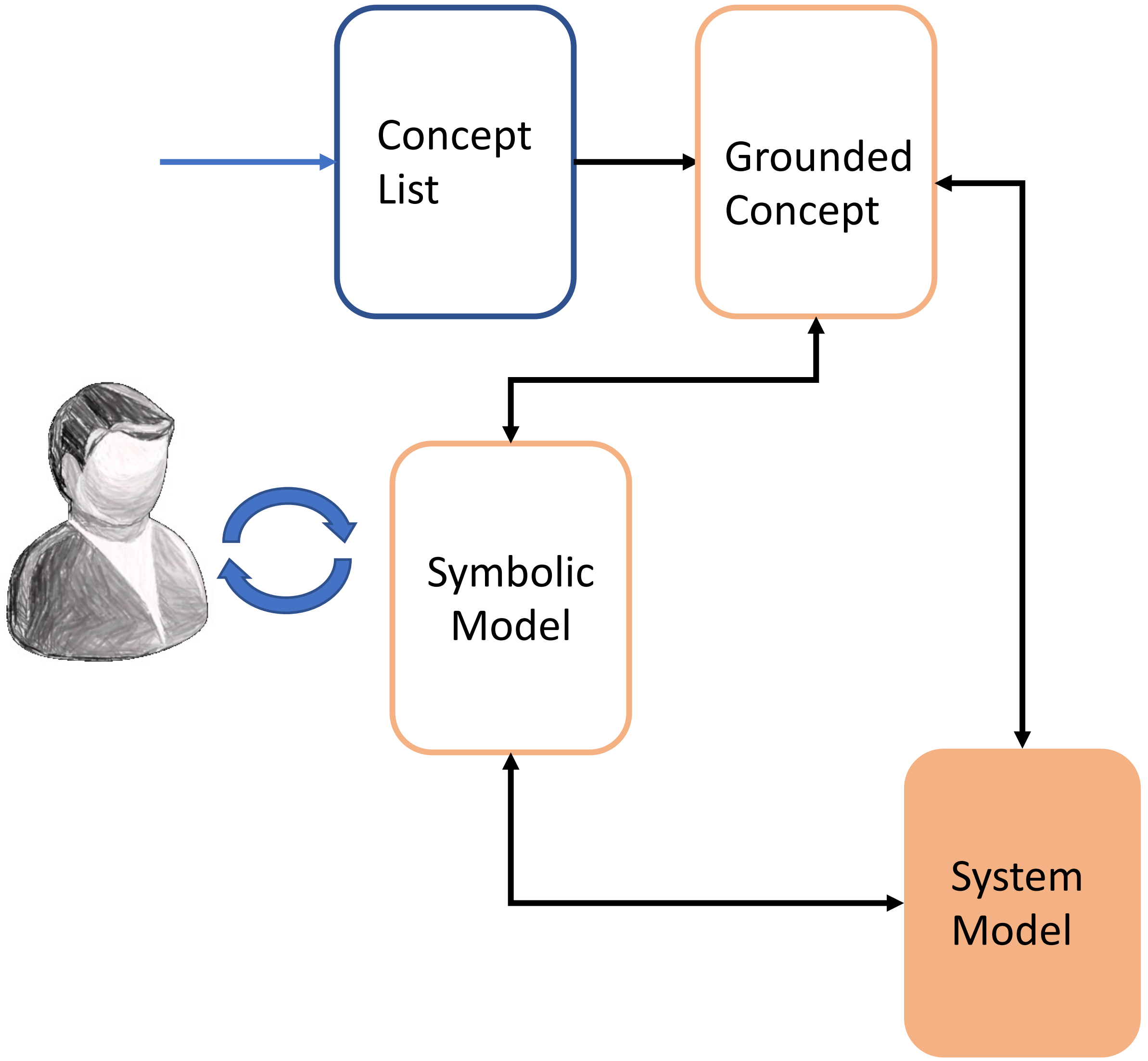
Challenge 2: Grounding Concept Set

- Next the concept set is grounded to learn the mapping between STST and individual concept as understood by the user
  - For specialized domains, this could mean the same concept may be grounded by different users in different ways
- One possible way to learn such grounded representations maybe to learn classifiers that identify whether a concept is present in an STST slice
  - User expected to provide positive and negative examples
- All learned groundings expected to be approximate and noisy
  - Any symbolic models learned should be capable of handling this level of noise



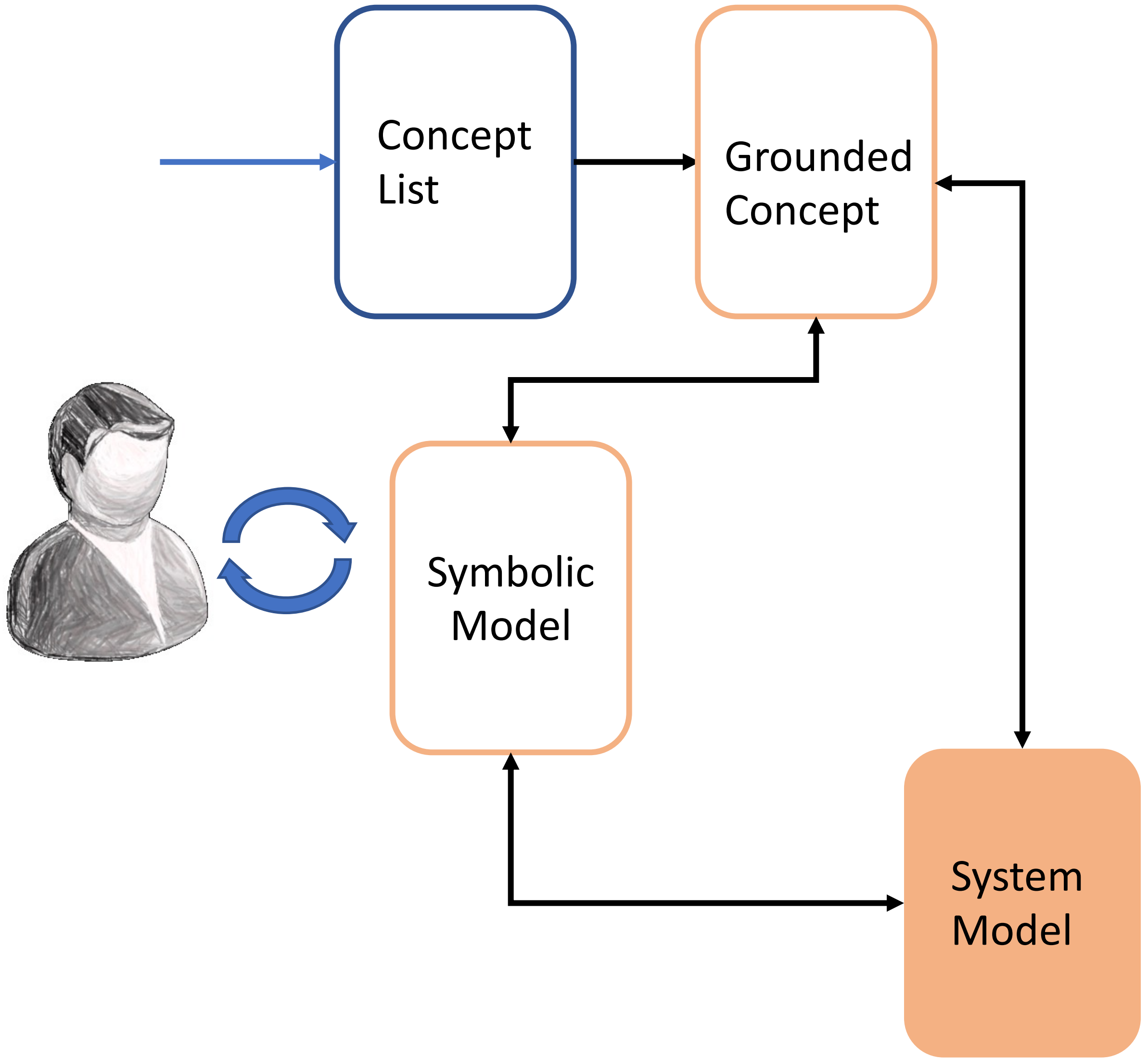
Challenge 3: Vocabulary Expansion

- Initial concept set bound to be incomplete with respect to its ability to represent the underlying model
- First challenge includes identifying vocabulary incompleteness
  - Requires the methods leveraging the symbolic models to be aware of the fact that the symbolic model may be incomplete and thus identify when the reasoning from the symbolic model may differ from the one obtained through the true model
- We have to engage in a process of ***vocabulary reconciliation*** to acquire missing yet necessary concepts for the task at hand



Challenge 3: Vocabulary Expansion (contd.)

- Two sources of incompleteness
  - User forgot to specify the concept
  - User's vocabulary does not include an equivalent concept
- The former requires the development of new techniques to that are able to efficiently query the human for previously unmentioned concepts
  - One could potentially use low-level explanations to guide the concepts the users may provide
- The latter requires the system to teach new concepts to humans
  - Early works in identifying concepts used by super-human AI systems like alphago presents interesting use-cases.



Summary

- Part 1: Why and how do humans exchange explanations? Do AI systems need to?
- Part 2: Using Mental Models for Explainable Behavior in the context of explicit knowledge tasks (think Task Planning)
  - The 3-model framework: $\mathcal{M}^R, \mathcal{M}^H, \mathcal{M}_h^R$
  - Explicability: *Conform to* $\mathcal{M}_h^R$
  - Explanation: Reconcile $\mathcal{M}_h^R$ to $\mathcal{M}^R$
  - Extensions: *Foils, Abstractions, Multiple Humans..*
- Part 3: Supporting explainable behavior even without shared vocabulary
  - Symbols as a *Lingua Franca* for Explainable and Advisable Human-AI Interaction
    - *Post hoc symbolic explanations of inscrutable reasoning*
    - *Accommodating symbolic advice into inscrutable systems*