

# Explanations & Explainability: Why do Humans care & How should AI Systems provide

Subbarao Kambhampati  
School of Computing & AI

**ASU**® Arizona State  
University

Research Funded in part by



Email: [rao@asu.edu](mailto:rao@asu.edu) Twitter: [@rao2z](https://twitter.com/rao2z) LinkedIn: [@Subbarao2z](https://www.linkedin.com/in/Subbarao2z)



# Explanations & Ex Why do Humans How should AI Syst

Subbarao Kambh  
School of Comput

Research Funded in part by



**ASU**® Arizona  
Univers

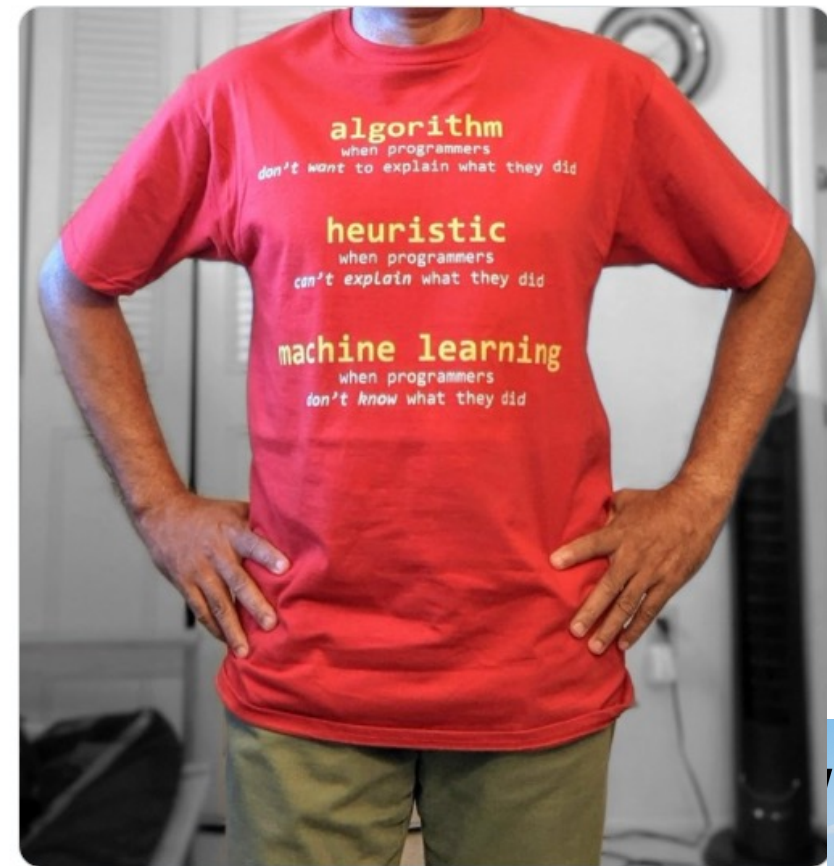
Email: [rao@asu.edu](mailto:rao@asu.edu) Twitter: [@rao2z](https://twitter.com/rao2z) LinkedIn: [@Subbarao2z](https://www.linkedin.com/in/Subbarao2z)



Subbarao Kambhampati (కంభంపాటి సుబ్బారావు)

@rao2z

Got my t-shirt and am now all set to teach about #Explainability in #AI (.. in terms even Intro #AI kids can understand..) 😎 #xai

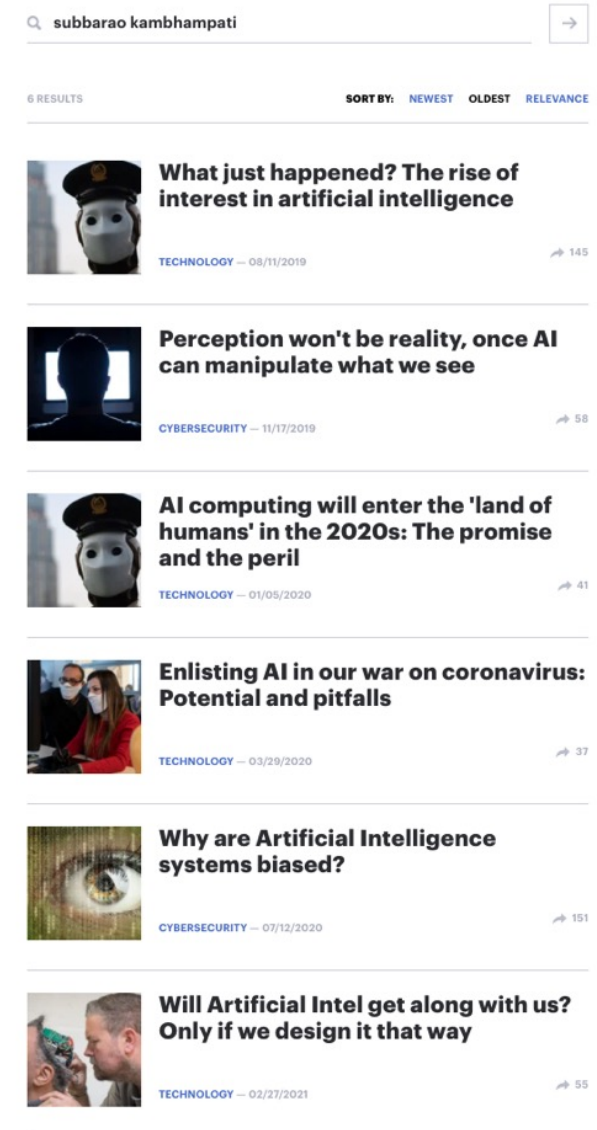


# About Me: Subbarao (Rao) Kambhampati

- Professor at School of Computing & AI at Arizona State University
- Former President of Association for Advancement of Artificial Intelligence (AAAI)
- Founding member of the Board of Directors of Partnership on AI
- Research in Human-Aware AI Systems; Explainable AI; Planning/Decision-Making
- Significant outreach/public dissemination on AI topics
  - Writes a column on The Hill









Twitter:  
@rao2z



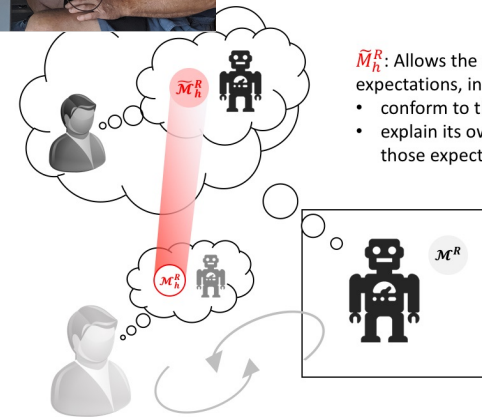
Q subbarao kambhampati

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**  
CYBERSECURITY — 11/17/2019 58
-  **AI computing will enter the 'land of humans' in the 2020s: The promise and the peril**  
TECHNOLOGY — 01/05/2020 41
-  **Enlisting AI in our war on coronavirus: Potential and pitfalls**  
TECHNOLOGY — 03/29/2020 37
-  **Why are Artificial Intelligence systems biased?**  
CYBERSECURITY — 07/12/2020 151
-  **Will Artificial Intel get along with us? Only if we design it that way**  
TECHNOLOGY — 02/27/2021 55

# 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

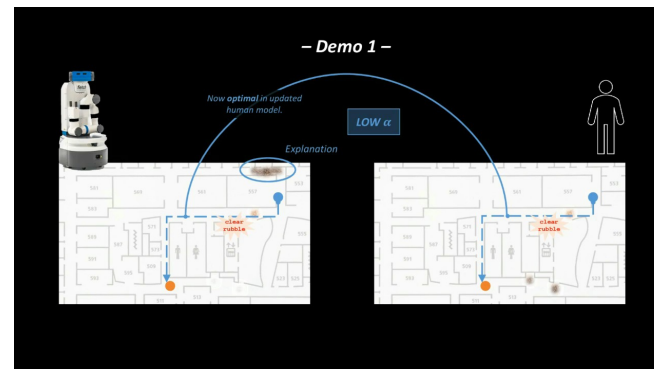
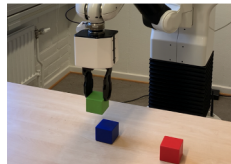


$\bar{M}_h^R$ : Allows the agent to anticipate human expectations, in order to

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

$M_h^H$  and  $\bar{M}_h^R$  are Expectations on Models  $M^H$  and  $M^R$

They don't have to be executable





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

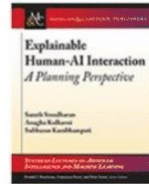
## Explainable Human-AI Interaction *A Planning Perspective*

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

Copyright © 2022 | 184 Pages

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.

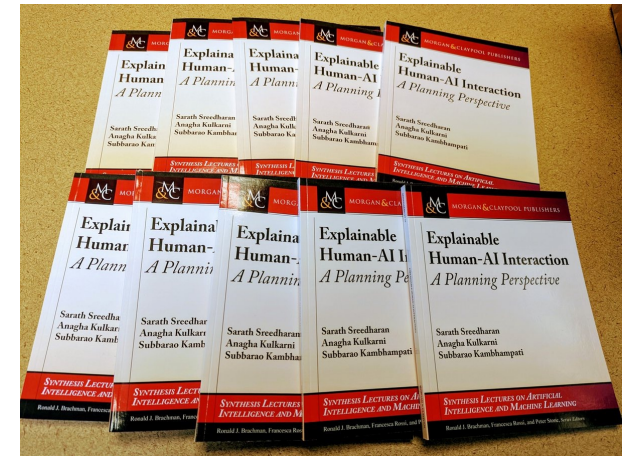
The authors draw from several years of research in their lab to discuss how an AI agent can use these mental models to either conform to human expectations or change those expectations through explanatory communication. While the focus of the book is on cooperative scenarios, it also covers how the same 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: State of AI + 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*



Blue Sky Talk  
AAAI 2022

## Symbols as a Lingua Franca for Bridging the Human-AI Chasm in Explainable & Advisable AI Systems

Subbarao Kambhampati  
(Joint with Sarath Sreedharan, Mudit Verma, Yantian Zha & Lin Guan)  
School of Computing & AI  
ASU  
Arizona State University

# Written vs. Learned Programs (Software)

## Traditional Programs

- Human programmers write the computer code
- The computer code executes and makes a decision
- Erroneous decisions can be traced directly back to the human programmer(s)

## AI & The Courts

(Briefing for NASEM Workshop on Emerging Areas of Science,  
Engineering & Medicine for the Courts)

Subbarao Kambhampati

**ASU** Arizona State  
University



 [rao@asu.edu](mailto:rao@asu.edu)  [@rao2z](https://twitter.com/rao2z)  [@subbarao2z](https://www.linkedin.com/in/subbarao2z)

## Learned Programs (AI)

- Human programmers writes general code schema to learn from data
- This general code is then trained on massive data corpora resulting in a “learned” program
- The learned program then executes and makes a decision
- Erroneous decisions are a complex combination of the general code schema and the training data
  - Quite often, the errors come from the training data
    - (E.g. An influential study showed that commercial gender recognition systems had high error rates for non-white-male subjects—mostly because they were trained on easily available data that happened to be unbalanced)



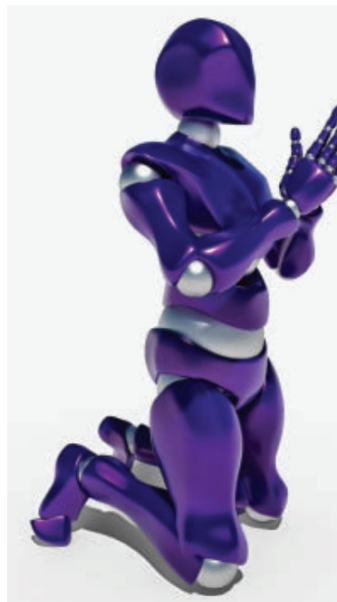
# Viewpoint

## 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.*

**I**N 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 accept things I cannot learn, data to learn the t I can, and wisdom to know the differenc**

sioned—for which we do have explicit

Twitter @rao2z

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/3aelxm1>), 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



subbarao kambhampati

6 RESULTS

SORT BY: NEWEST OLDEST RELEVANCE



**What just happened? The rise of interest in artificial intelligence**

TECHNOLOGY — 08/11/2019

145

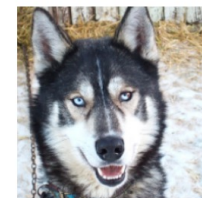


**Perception won't be reality, once AI can manipulate what we see**



# 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.
  - Explanation here helps reconcile the expectations
- 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)
  - 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



(a) Husky classified as wolf



(b) Explanation

**Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.**

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

# 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 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)

Explanations in Law are often meant to be symbolic (explicit)



# 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



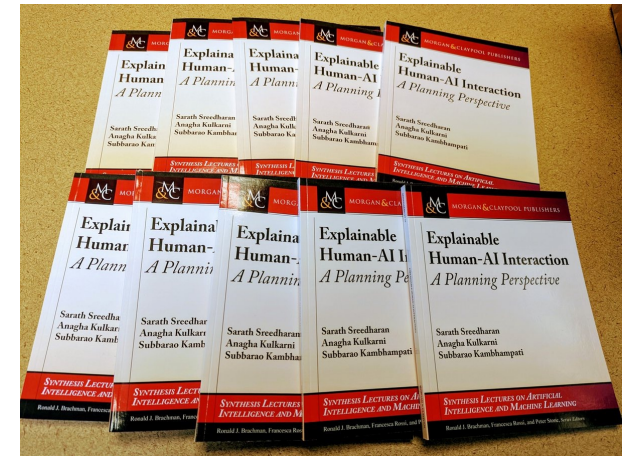
# Requirements on Explanations

- **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/ChatGPT

Contestability

# Talk Overview

- Part 1: State of AI + 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*



Blue Sky Talk  
AAAI 2022

## Symbols as a Lingua Franca for Bridging the Human-AI Chasm in Explainable & Advisable AI Systems

Subbarao Kambhampati  
(Joint with Sarath Sreedharan, Mudit Verma, Yantian Zha & Lin Guan)  
School of Computing & AI  
ASU  
Arizona State University

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

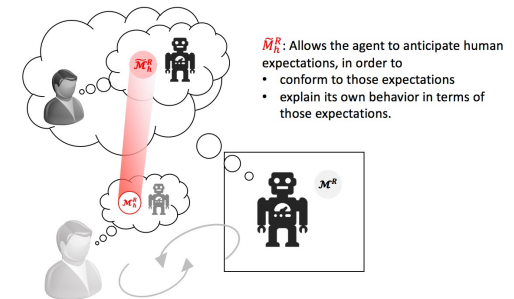
## Managing Mental Models

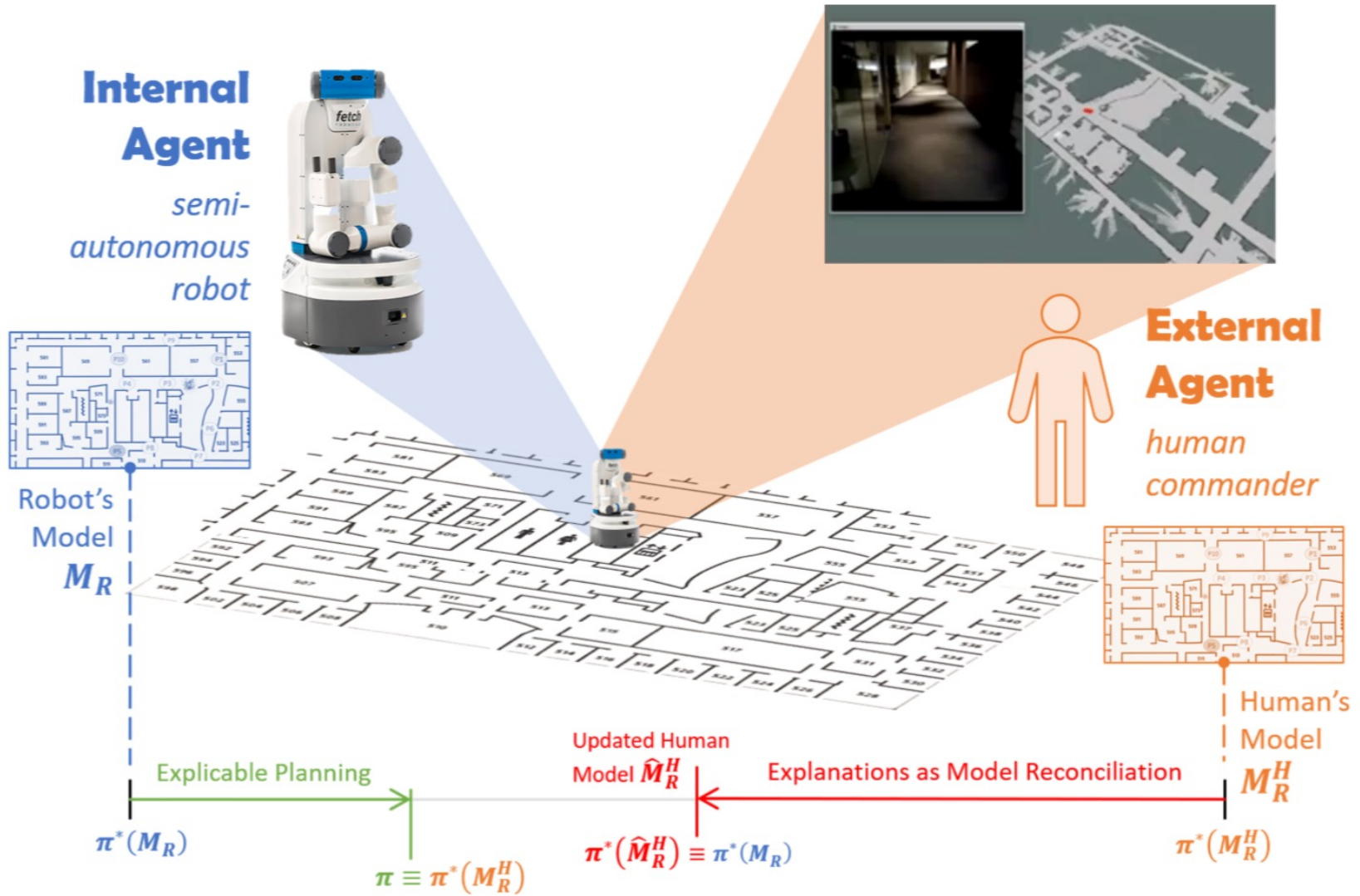




# Model differences with human in the loop

- 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





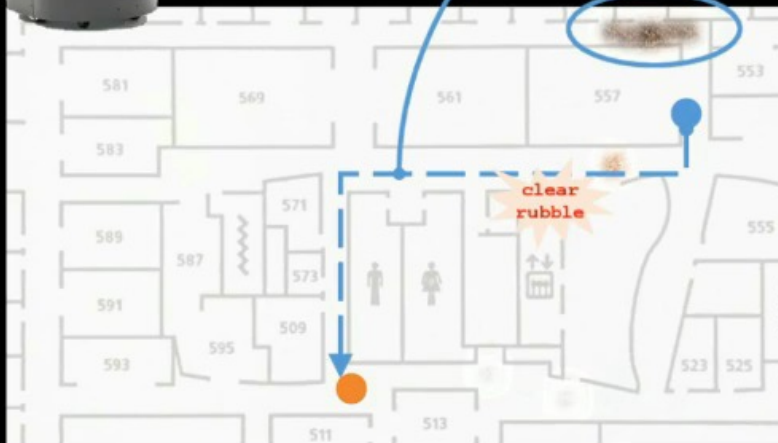
– Demo 1 –



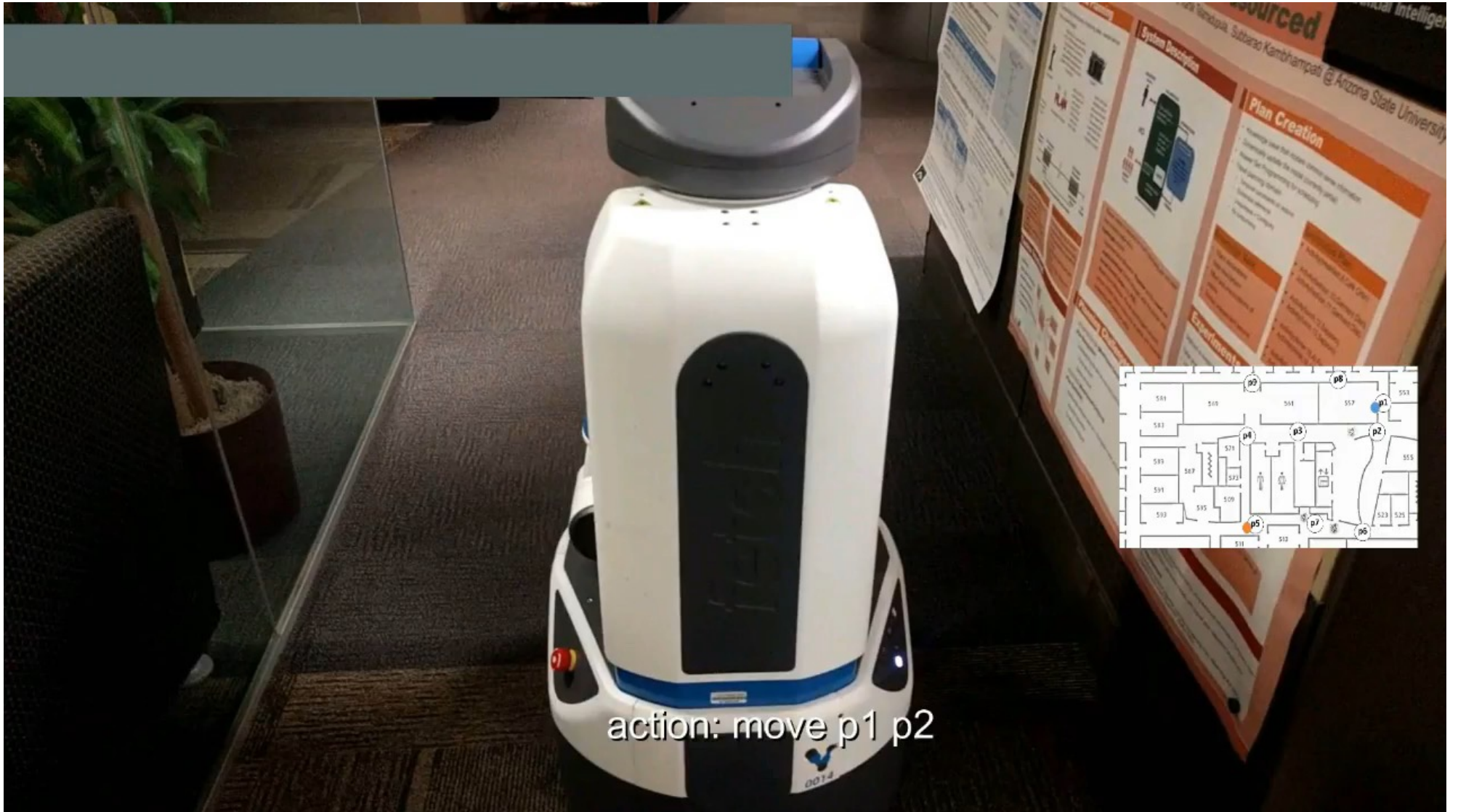
Now *optimal* in updated human model.

LOW  $\alpha$

Explanation







action: move p1 p2

# Model Space Search for Model Reconciliation

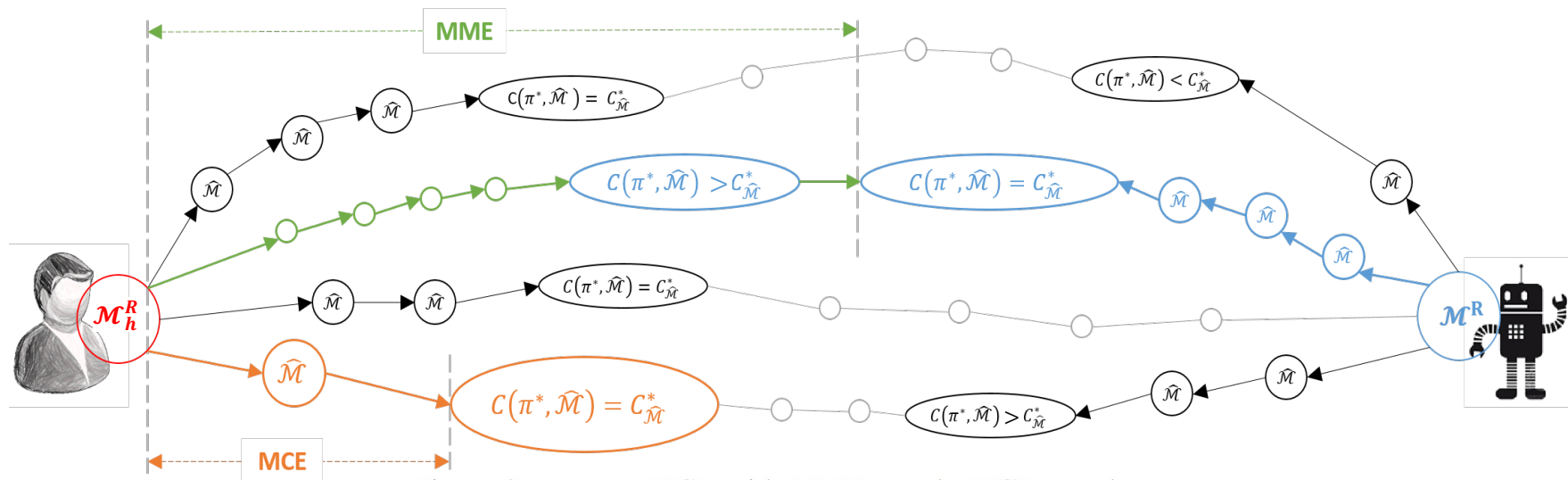
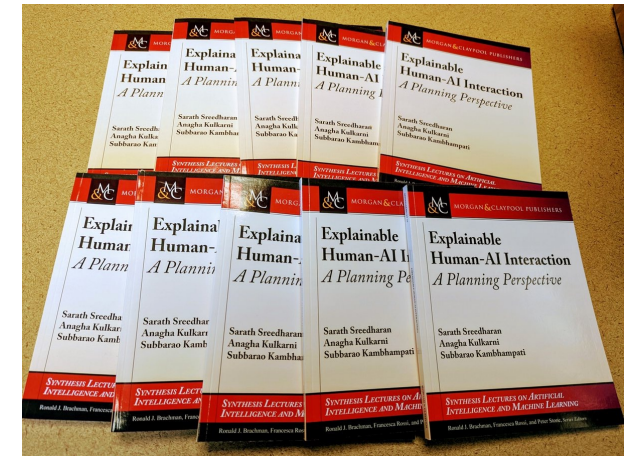


Figure 3 contrasts MCE with MME search. MCE search starts from  $\mathcal{M}^H$ , computes updates  $\hat{\mathcal{M}}$  towards  $\mathcal{M}^R$  and returns the first node (indicated in orange) where  $C(\pi^*, \hat{\mathcal{M}}) = C_{\hat{\mathcal{M}}}^*$ . MME search starts from  $\mathcal{M}^R$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

# Talk Overview

- Part 1: State of AI + 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*



Blue Sky Talk  
AAAI 2022

## Symbols as a Lingua Franca for Bridging the Human-AI Chasm in Explainable & Advisable AI Systems

Subbarao Kambhampati  
(Joint with Sarath Sreedharan, Mudit Verma, Yantian Zha & Lin Guan)  
School of Computing & AI  
ASU  
Arizona State University



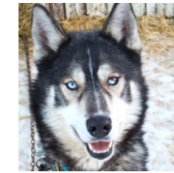
# Explanations in the absence of shared vocabulary

- 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!

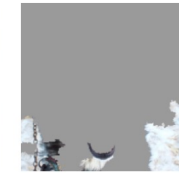


(a) Original Image

Figure 4: Explaining a lighting positive pixels. ( $p = 0.24$ ) and “Labrad



(a) Husky classified as wolf



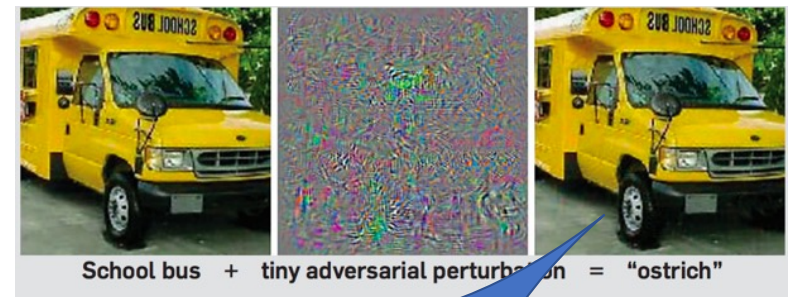
(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.

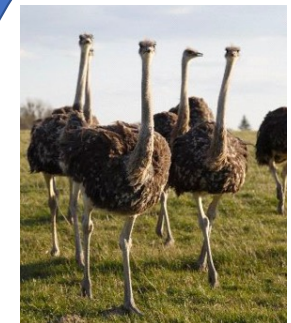


Explaining Labrador

on network, high-“Acoustic guitar”



Please point to the “ostrich” parts



# 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..)

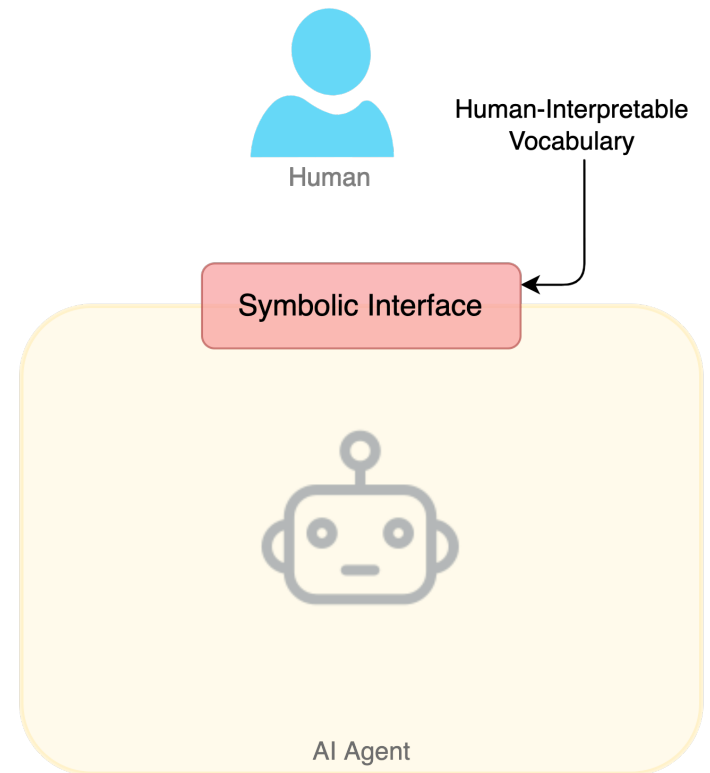
- Typically, pointing explanations are used for tacit knowledge tasks, and symbolic ones for explicit knowledge tasks.
  - However, over time, we tend to develop symbolic vocabulary for exchanging explanations even for tacit knowledge tasks.
    - Consider, for example, Pick-and-Roll in Basketball..
- 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)

Explanations in Law are often meant to be symbolic (explicit)



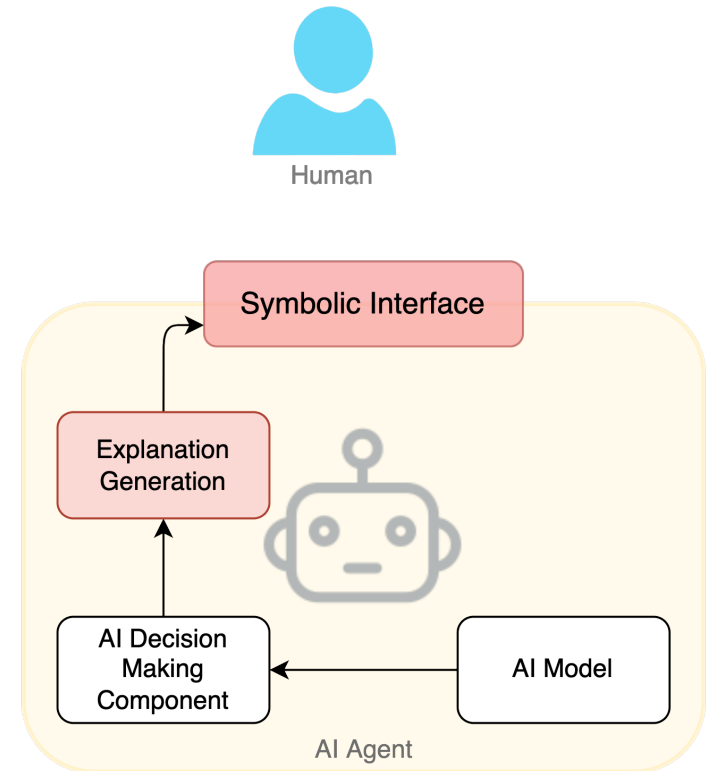
# Use case for the Symbolic Layer

- 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

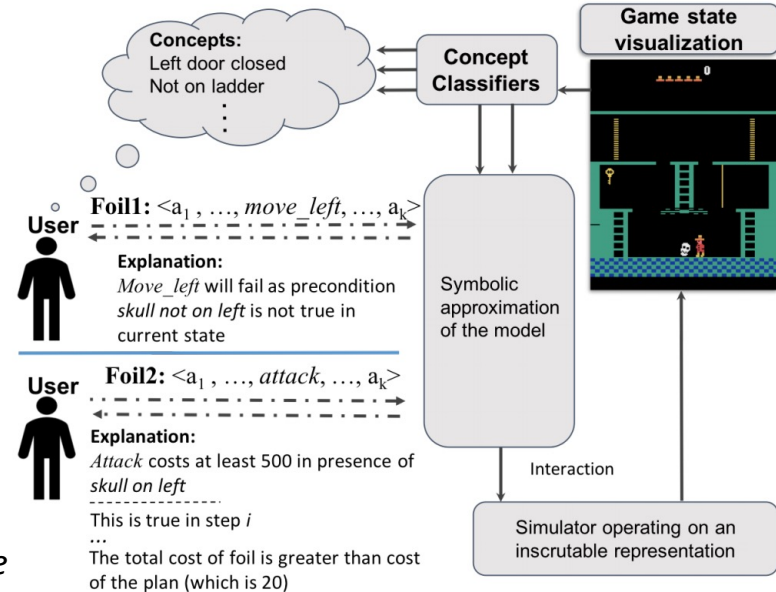
User raises a foil – i.e., an alternate plan – A model component learned to refute the foil

The foil fails at any point

*Identify the missing preconditions*

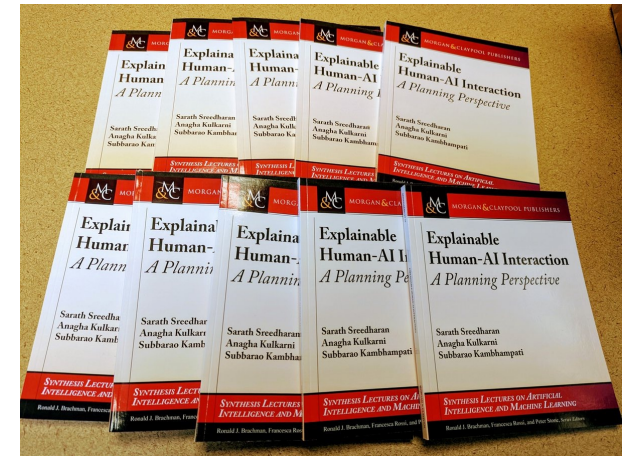
The foil is costlier than the original plan

*Identify an abstract version of the cost function*



# Talk Overview

- Part 1: State of AI + 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*



Blue Sky Talk  
AAAI 2022

## Symbols as a Lingua Franca for Bridging the Human-AI Chasm in Explainable & Advisable AI Systems

Subbarao Kambhampati  
(Joint with Sarath Sreedharan, Mudit Verma, Yantian Zha & Lin Guan)  
School of Computing & AI  
ASU  
Arizona State University

# Welcome to the AIES 2023 Conference Site



## AAAI / ACM conference on **ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY**

### **Call for papers:**

To be announced. Please stay tuned

### **Important dates:**

Conference dates: August, 2023 (tentative)

**AIES 2023 will be held in Montreal.**

<https://www.aies-conference.com/2023/>