rao@asu.edu @rao2z @subbarao2z

Challenges of Human-Aware AI Systems

Subbarao Kambhampati

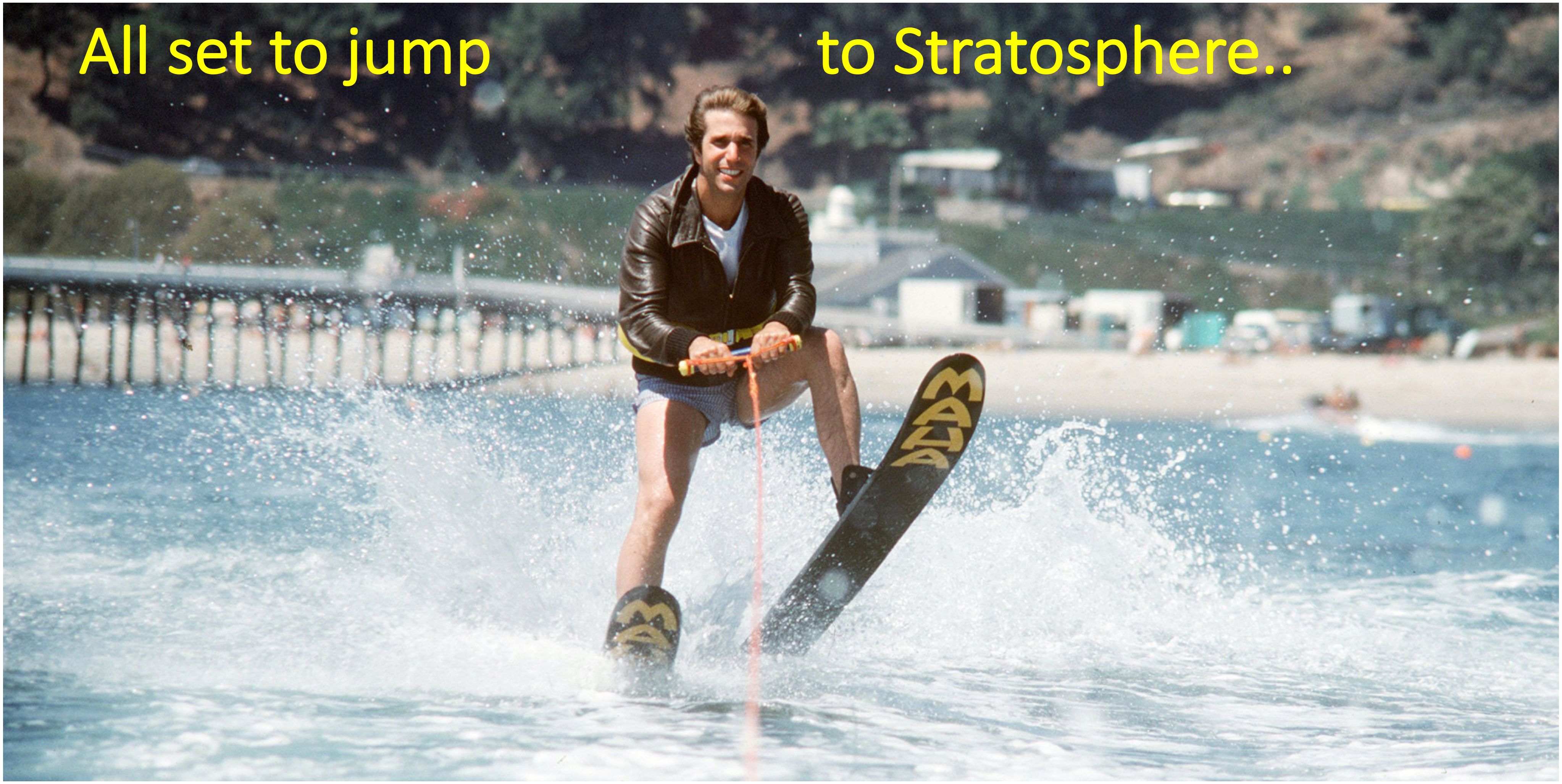
#MAIGA

- AI is the new electricity (Ng)
- AI is bigger than FIRE & ELECTRICITY (Pichai)
- AI is GOD (Levandowski)

- AI is bigger threat than North Korea. [..] AI will start the third world war (Musk)
- AI could be the "worst event in the history of civilization" (Hawking)

AI is highly likely to destroy humans (Musk)

The new ~~Colgate~~ Smart Electronic Toothbrush provides real-time feedback to improve brushing habits and help prevent problems before they start. Designed with the help of dentists, the brush features real-time sensors and artificial intelligence algorithms to detect brushing effectiveness in 16 zones of the mouth.


All set to jump
to Stratosphere..

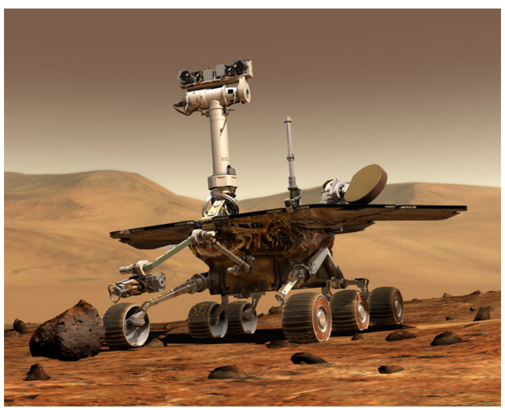

LIVE
breakyourownnews.com
FAKE
NEWS
BREAKING NEWS
AI HELPS OLD LADY CROSS STREET!
16:43
AI PLAYS WITH KIDS, COOKS FOOD, AND HANGS AROUND SANS DRAMA

Objective of this talk..

- Why isn't human-aware AI all over the place already?
- Why we should pursue it? (Hint: It broadens the scope & promise of AI)
- Research Challenges in HAAI (Case Study: Our research on Human-aware Planning & Decision Making)
- Long term issues (Trust); Ethical Dilemmas

AI's Curious Ambivalence to humans..

- Our systems seem happiest
  - either far away from humans
  - or in an adversarial stance with humans



You want to help humanity,
it is the people that you just can't stand…

What happened to Co-existence?

- Whither McCarthy's advice taker?
- ..or Janet Kolodner's house wife?
- …or even Dave's HAL?
  - (with hopefully a less sinister voice)

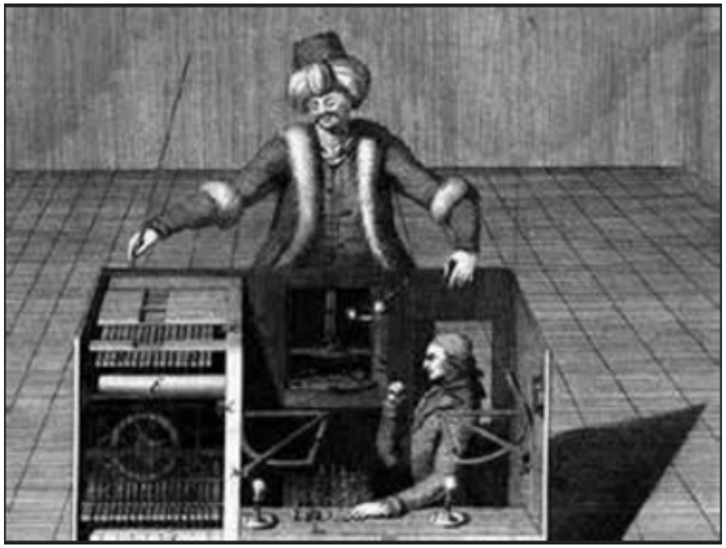
HAAI

Human-aware AI

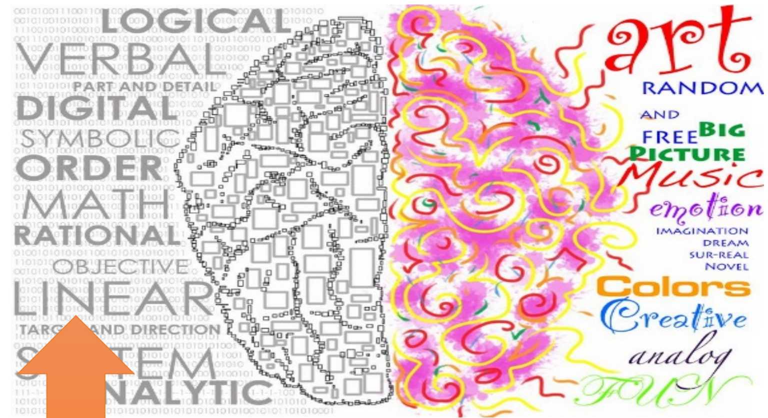
But isn't this cheating?

- Doesn't putting human in the loop dilute the AI problem?
- Won't it be cheating?
  - Like the original Mechanical Turk.. or the more recent Mechanical Saud..

  - (or the early mixed-initiative planners, that had humans helping an automated planner by manipulating its search queue)



The Many Intelligences..



Humans

- Perceptual & Manipulation intelligence that seem to come naturally to us
  - Image recognition; hand-eye coordination
  - Largely tacit
- Emotional Intelligence
  - Showing & recognizing emotional responses
- Social Intelligence
  - Requires a "theory of mind"
- Cognitive/reasoning tasks
  - That seem to be what we get tested in in SAT etc.
  - (More declarative..)

AIs

HAAI is needed Everywhere..

- There are of course areas where humans are *sine qua non* (..and received attention)
    - Intelligent Tutoring Systems
        - Pioneering work by researchers such as Kurt van Lehn
    - Social Robotics
        - Pioneering work by researchers such as Cynthia Brazeal, Brian Scassallati

- ..but those are not all! we need HAAI in even quotidian situations
    - Assistance
        - Human-aware digital personal assistants
        - Human-aware office/hospital assistants
    - Teaming
        - Elbow-to-Elbow (Factory Floor)
        - Remote/Cognitive (Search & Rescue; Mixed-initiative/cooperative planning/decision-making)

- Increasingly, HCI will Human-AI Interaction

AI Magazine Volume 17 Number 2 (1996) (© AAAI)

AAAI–94 Presidential Address

Collaborative Systems

Barbara J. Grosz

■ The construction of computer systems that are intelligent, collaborative problem-solving partners is an important goal for both the science of AI and its application. From the scientific perspective, the development of theories and mechanisms to enable building collaborative systems presents exciting research challenges across AI subfields. From the applications perspective, the capability to collaborate with users and other systems is essential if large-scale information systems of the future are to assist users in finding the information they need and solving the problems they have. In this address, it is argued that collaboration must be designed into systems from the start; it cannot be patched on. Key features of collaborative activity are described, the scientific base provided by recent AI research is discussed, and several of the research challenges posed by collaboration are presented. It is further argued that research on, and the development of, collaborative systems should itself be a collaborative endeavor—within AI, across subfields of computer science, and with researchers in other fields.

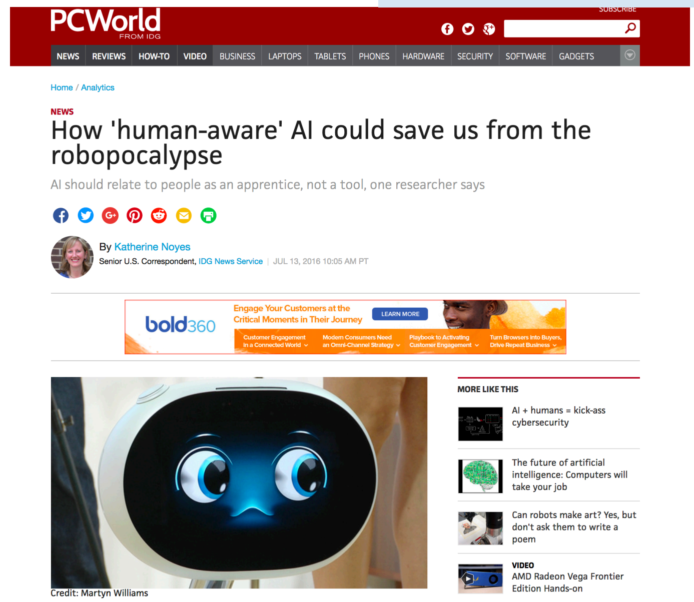
AI has always pushed forward on the frontiers of computer science. Our efforts to understand intelligent behavior and the ways in which it could be embodied in computer systems have led both to a richer scientific understanding of various aspects of intelligence and to the development of smarter computer systems. In his keynote address at AAAI-94, Raj Reddy standing of collaborative systems and the development of the foundations—the representations, theories, computational models and processes—needed to construct computer systems that are intelligent collaborative partners in solving their users' problems. In doing so, I follow the precedent set by Allen Newell in his 1980 Presidential Address (Newell 1981, p. 1) of focusing on the state of the science rather than the state of the society. I also follow a more recent precedent, that set by Daniel Bobrow in his 1990 Presidential address (Bobrow 1991, p. 65), namely, examining the issues to be faced in moving beyond what he called the "isolation assumptions" of much of AI to the design and analysis of systems of multiple agents interacting with each other and the world. I concur with his claim that a significant challenge for AI in the 1990s is "to build AI systems that can interact productively with each other, with humans, and with the physical world" (p. 65). I will argue further, however, that there is much to be gained by looking in particular at one kind of group behavior, collaboration.

My reasons for focusing on collaborative systems are two-fold. First, and most important in this setting, the development of the underlying theories and formalizations that are needed to build collaborative systems as well as the construction of such systems raises interesting questions and presents intellectual challenges across AI subfields. Sec-

This talk was presented at the American Association for Artificial Intelligence's National Conference on Artificial Intelligence, 3 August 1994, in Seattle, Washington

Why intentionally design a dystopian future and spend time being paranoid about it?

AAAI-18 Special Track on Human-AI Collaboration!



8.2 Recommendations

JASON offers the following recommendations to DoD senior leadersh

1. DoD should both track (via a knowledgeable cadre) and invest portfolio) the most dynamic and rapidly advancing areas of AI means limited to DL.

55

JASON Briefing on "The Path t General AI goes through Aware AI"; June 2016

2. DoD should support the development of a discipline of AI eng progress of the field through Shaw's "craft" and (empirical) "c particular focus should be advancing the "illities" in support of
3. DoD's portfolio in AGI should be modest and recognize that it advancing area of AI. The field of human augmentation via AI and deserves significant DoD support.
4. DoD should support the curation and labeling, for research, of large data sets. Wherever possible, operational data should be s use in support of AI for DoD-unique missions.
5. DoD should create and provide centralized resources for its int researchers (MOSIS-like), including labeled data sets and acce training platforms.
6. DoD should survey the mission space of embedded devices for applications of AI, and should consider investing in special-pu support AI inference in embedded devices for DoD missions if identified.

Seeking new algorithms for human-aware AI

Over the years, AI algorithms have become able to solve problems of increasing complexity. However, there is a gap between the capabilities of these algorithms and the usability of these systems by humans. *Human-aware* intelligent systems are needed that can interact intuitively with users and enable seamless machine-human collaborations. Intuitive interactions include shallow interactions, such as when a user discards an option recommended by the system; model-based approaches that take into account the users' past actions; or even deep models of user intent that are based upon accurate human cognitive models. Interruption models must be developed that allow an intelligent system to interrupt the human only when necessary and appropriate. Intelligent systems should also have the ability to augment human cognition, knowing which information to retrieve when the user needs it, even when they have not prompted the system explicitly for that information. Future intelligent systems must be able to account for human social norms and act accordingly. Intelligent systems can more effectively work with humans if they possess some degree of emotional intelligence, so that they can recognize their users' emotions and respond appropriately. An additional research goal is to go beyond interactions of one human and one machine, toward a "systems-of-systems", that is, teams composed of multiple machines interacting with multiple humans.

Human-AI system interactions have a wide range of objectives. AI systems need the ability to represent a multitude of goals, actions that they can take to reach those goals, constraints on those actions, and other factors, as well as easily adapt to modifications in the goals. In addition, humans and AI systems

NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

must share common goals and have a mutual understanding of them and relevant aspects of their current states. Further investigation is needed to generalize these facets of human-AI systems to develop systems that require less human engineering.

THEMATIC PILLARS

1. Safety-critical AI

Advances in AI have the potential to improve outcomes, enhance quality, and reduce costs in such safety-critical areas as healthcare and transportation. Effective and careful applications of pattern recognition, automated decision making, and robotic systems show promise for enhancing the quality of life and preventing thousands of needless deaths.

However, where AI tools are used to

2. Fair, Transparent, and Accountable AI

AI has the potential to provide societal value by recognizing patterns and drawing inferences from large amounts of data. Data can be harnessed to develop useful diagnostic systems and recommendation engines, and to support people in making breakthroughs in such areas as biomedicine, public health, safety, criminal justice, education, and sustainability.

3.Collaborations between people and AI systems

A promising area of AI is the design of systems that augment the perception, cognition, and problem-solving abilities of people.

Examples include the use of AI technologies to help physicians make more timely and accurate diagnoses and assistance provided to drivers of cars to help them to avoid dangerous situations and crashes.

4. AI, labor and the economy

AI advances will undoubtedly have multiple influences on the distribution of jobs and nature of work. While advances promise to inject great value into the economy, they can also be the source of disruptions as new kinds of work are created and other types of work become less needed due to automation.

Discussions are rising on the best approaches to minimizing potential disruptions, making sure that the fruits of AI advances are widely shared, and competition and innovation is encouraged and not stifled. We seek to study and understand best paths forward, and play a role in this discussion.

5. Social and societal influences of AI

AI advances will touch people and society in numerous ways, including potential influences on privacy, democracy, criminal justice, and human rights. For example, while technologies that personalize information and that support people with recommendations can provide people with valuable assistance, they could also inadvertently or deliberately manipulate and influence opinions.

We seek to promote thoughtful collaboration and open dialogue about the potential subtle and salient influences of AI on people and society.

6. AI for social good

AI offers great potential for promoting the public good, for example in the realms of education, housing, public health, and sustainability. We see great value in collaborating with public and private organizations, including academia, scientific societies, NGOs, social entrepreneurs, and interested private citizens to promote discussions and catalyze efforts to address society's most pressing challenges.

Some of these projects may address deep societal challenges and will be moonshots - ambitious big bets that could have far-reaching impacts. Others may be creative ideas that could quickly produce positive results by harnessing AI advances.

7. Special initiatives

Beyond the specified thematic pillars, we also seek to convene and support projects that resonate with the tenets of our organization. We are particularly interested in supporting people and organizations that can benefit from the Partnership's diverse range of Partners.

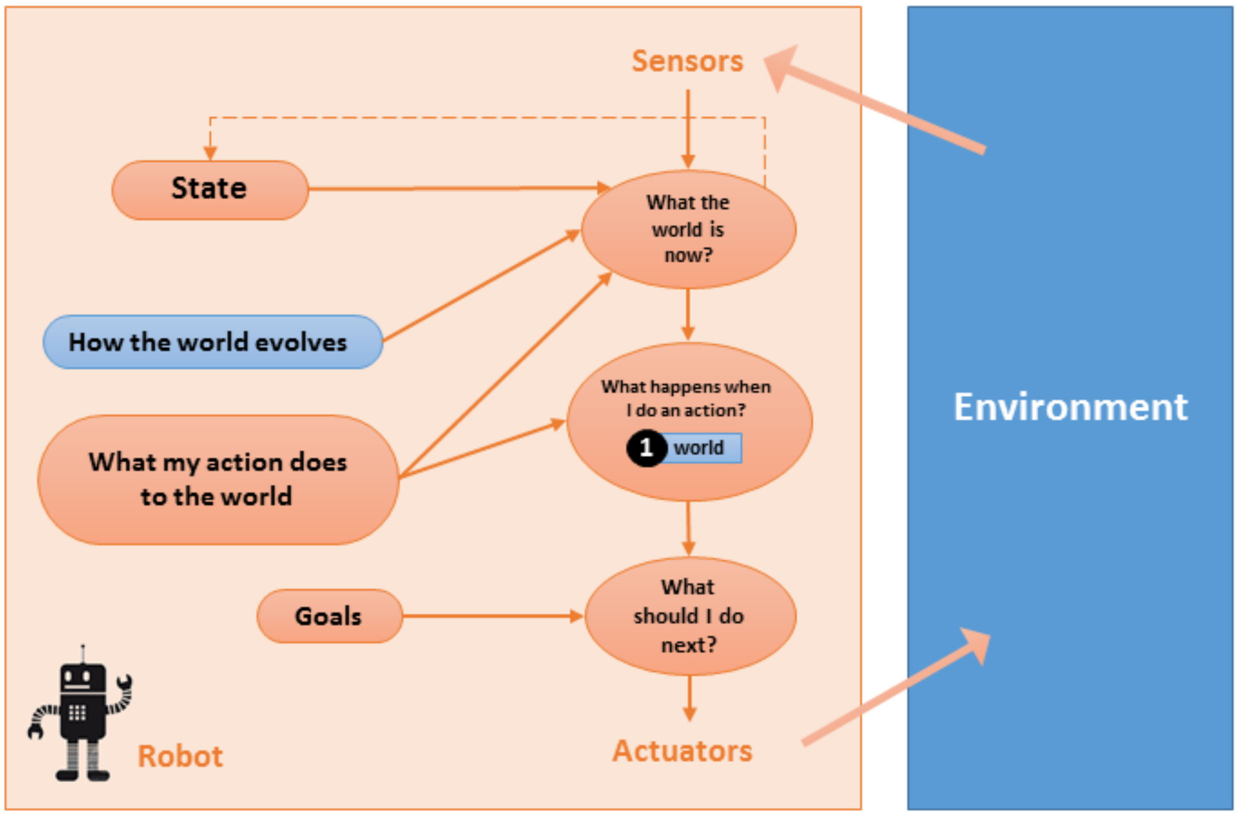
We are open-minded about the forms that these efforts will take.

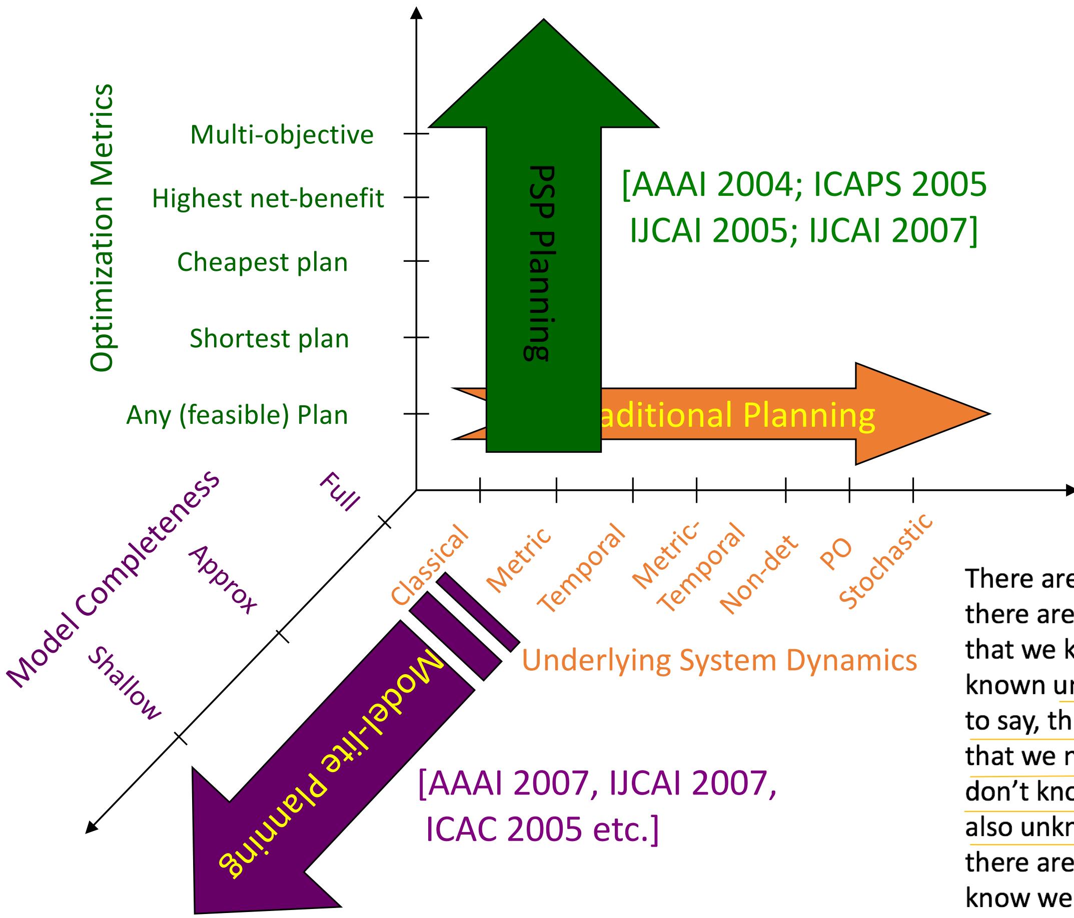
Objective of this talk..

- Why isn't human-aware AI all over the place already?
- Why we should pursue it? (Hint: It broadens the scope & promise of AI)
- Research Challenges in HAAI (Case Study: Our research on Human-aware Planning & Decision Making)
- Long term issues (Trust); Ethical Dilemmas


fetch
Thanks also to
Behzad Kamgar-Parsi
Benjamin Knott
Marc Steinberg
Jeffrey Morrison

Architecture of an Intelligent Agent

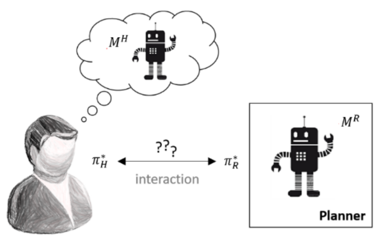


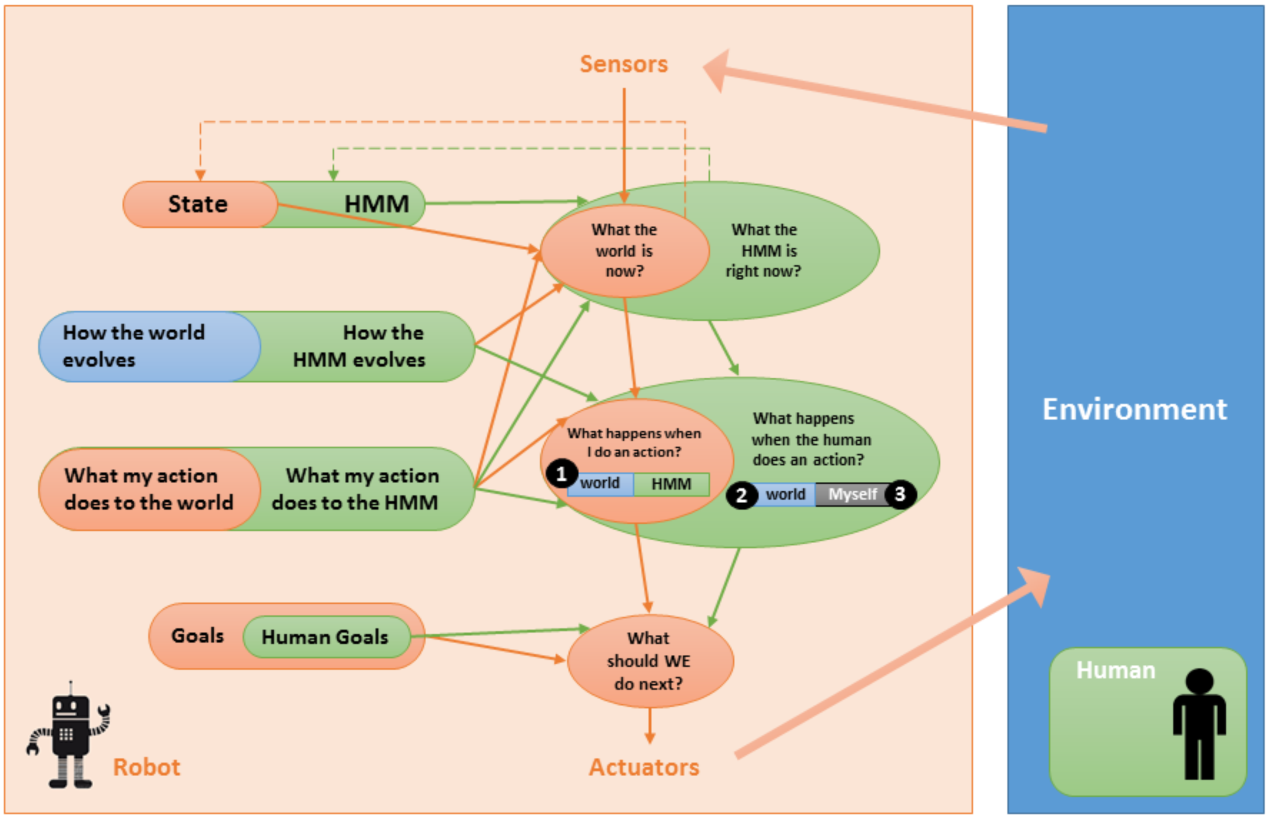


There are known knowns; there are things we know that we know. There are known unknowns; that is to say, there are things that we now know we don't know. But there are also unknown unknowns; there are things we do not know we don't know.

ASIDE: Interesting connections with Dietterich's 2016 address.

Architecture of an Intelligent Agent teaming with a human





HMM= Human Mental Model

Computer Science > Artificial Intelligence

AI Challenges in Human-Robot Cognitive Teaming

Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, Yu Zhang

(Submitted on 15 Jul 2017)

Among the many anticipated roles for robots in future is that of being a human teammate. Aside from all the technological hurdles that have to be overcome on the hardware and control sides to make robots fit for work with humans, the added complication here is that humans have many conscious and subconscious expectations of their teammates -- indeed, teaming is mostly a cognitive rather than physical coordination activity. This focus on cognitive coordination, however, introduces new challenges for the robotics community that require fundamental changes to the traditional view of autonomous agents.

In this paper, we provide an analysis of the differences between traditional autonomous robots and robots that team with humans, identifying the necessary teaming capabilities that are largely missing from current robotic systems. We then focus on the important challenges that are unique and of particular importance to human-robot teaming, especially from the point of view of the deliberative process of the autonomous agent, and sketch potential ways to address them.

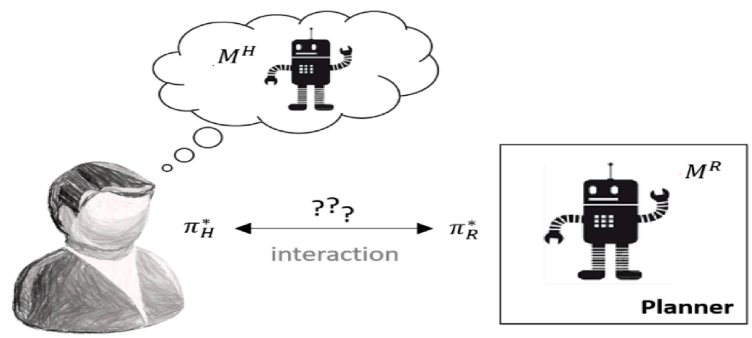
Subjects: **Artificial Intelligence (cs.AI)**

Cite as: arXiv:1707.04775 [cs.AI]

(or arXiv:1707.04775v1 [cs.AI] for this version)

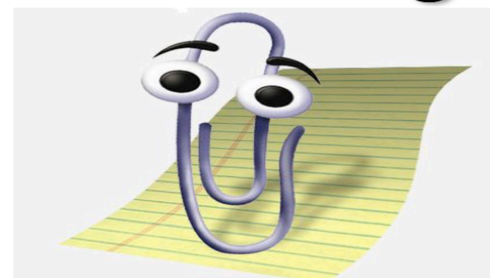
HAAI Challenges [With Focus on Planning & Decision Making]

- The primary challenge is modeling & reasoning with human mental models. Specifically:
- Modeling & Managing the human's mental state
  - Intention recognition; Intention projection
- Modeling & Managing the human's model of the AI System
  - Critical for the system to show (i) explicable behavior (ii) provide explanations of its decisions (iii) balance explicability & explanations



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

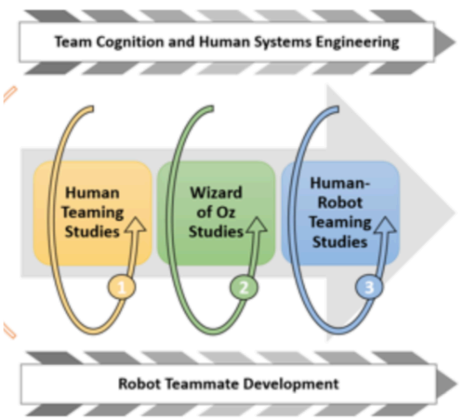
Proactive Help Can be Disconcerting!



Our solution: Interdisciplinary collaboration.

Solution: Interdisciplinary Collaboration

- Long-term collaboration with Prof. Nancy Cooke
  - Past President of Human Factors and Ergonomics Society
  - Expertise in human-human teaming; team performance etc.



Prof. Nancy Cooke;
Past President of Human Factors Society

Human-human Teaming Analysis in Urban Search and Rescue

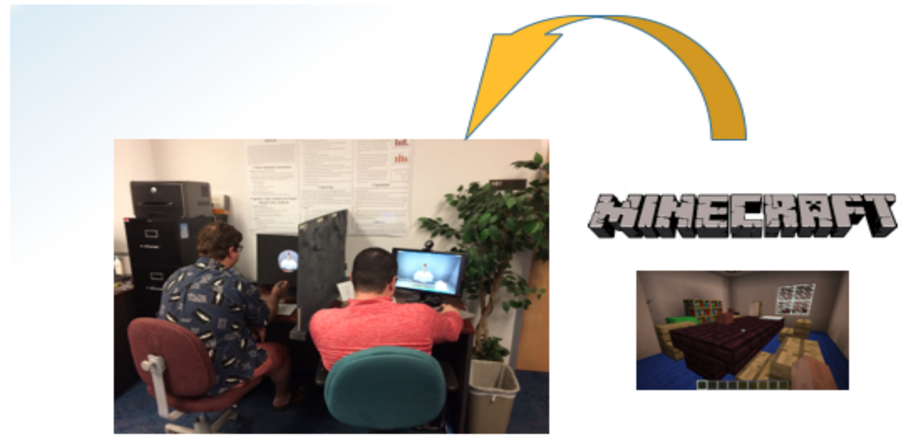
Simulated search task (Minecraft) with human playing role of USAR robot

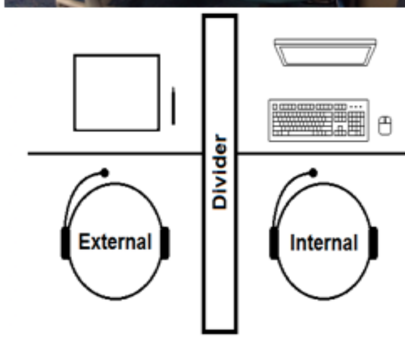
- 20 internal/external dyads tested
- Conditions of autonomous/intelligent or remotely controlled robot
- Differences in SA, performance, and communications

Urban Search and Rescue Task

- Simulated search task (Minecraft) with human playing role of USAR robot
  - 50 internal/ external dyads
  - A 2x2 design

	Mental Models	
Communication	Natural Language & Shared Models	Natural Language & Restricted Models
	Limited Language & Shared Models	Limited Language & Restricted Models



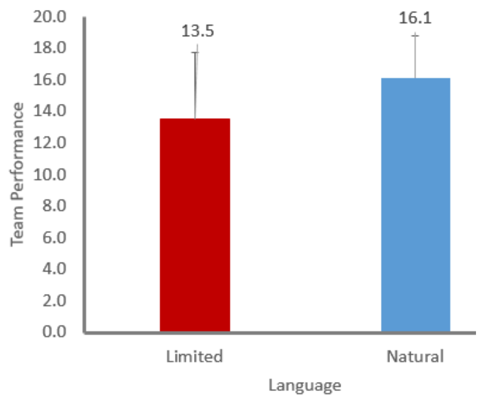


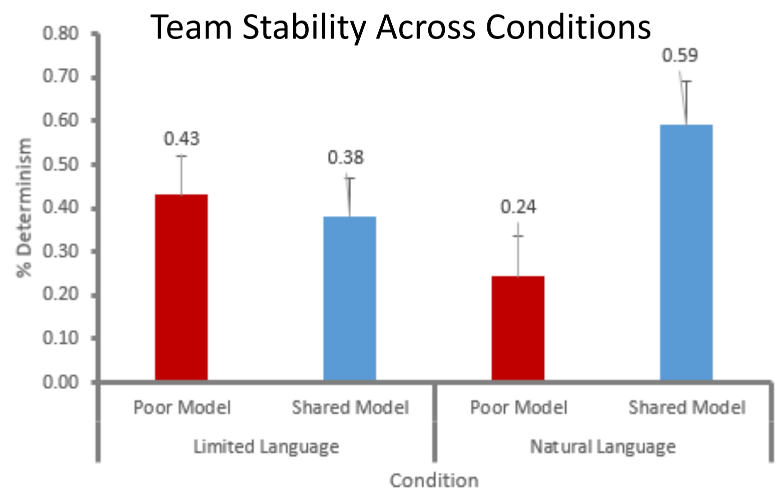
Participant positioning for experiment.

➢ Measures
- Team Performance
- Team Verbal Behaviors
- Team Situation Awareness
- NASA TLX Workload
- Team Synchrony

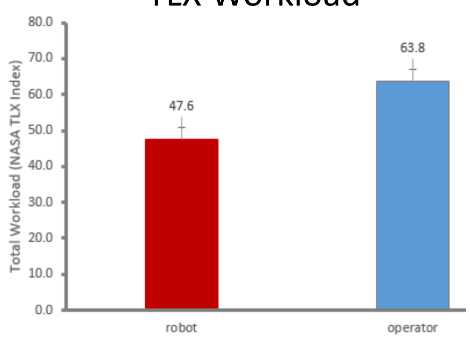
Sample Results

Language and Performance





Participant Role and NASA TLX Workload



Conclusions:

- Restricted language on part of “robot” hurt team performance
- Dyads using natural language and shared mental models had more stable behavior than other dyads
- When “robot” unaware of operator’s challenges, operator perceives higher workload than when “robot” is aware

Teaming Requires Modeling the Human

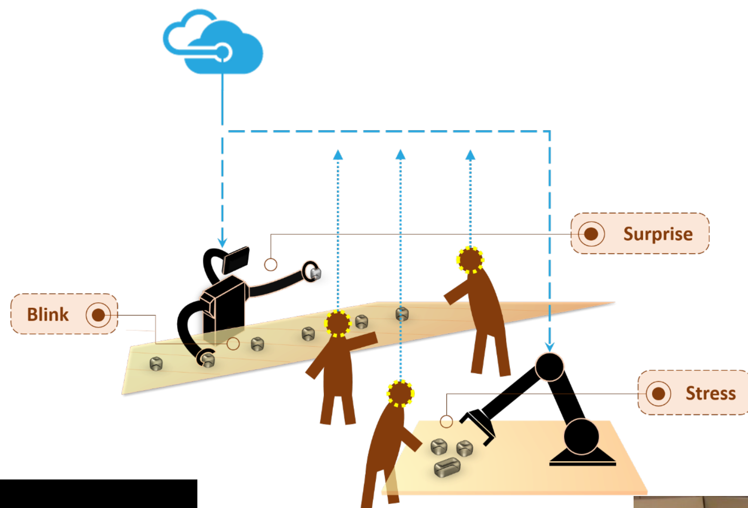
- "Theory of Mind"
- Intention recognition
  - What are they trying to achieve?
  - Allows for proactive support
    - [AAMAS 2016; HRI 2015; IROS 2015]
- Intention projection
  - Give them heads-up on what you are doing
    - [IROS 2015]

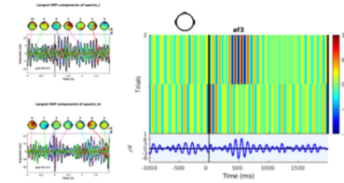
Planning with Resource Conflicts in Human-Robot Cohabitation
Tathagata Chakraborti, Yu Zhang, Subbarao Kambhampati.
AAMAS 2016.

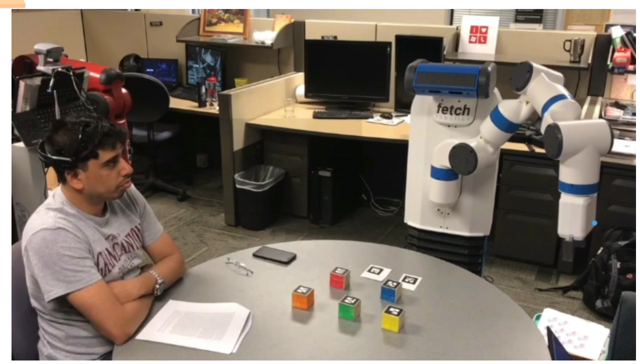
A Human Factors Analysis of Proactive Assistance in Human-robot Teaming.
Yu (Tony) Zhang, Vignesh Narayanan, Tathagata Chakraborti & Subbarao Kambhampati.
IROS 2015.

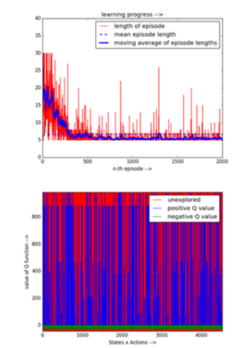
Planning for Serendipity.
Tathagata Chakraborti, Gordon Briggs, Kartik Talamadupula, Yu Zhang, Matthias Scheutz, David Smith and Subbarao Kambhampati
IROS 2015

Intention Recognition with Emotive

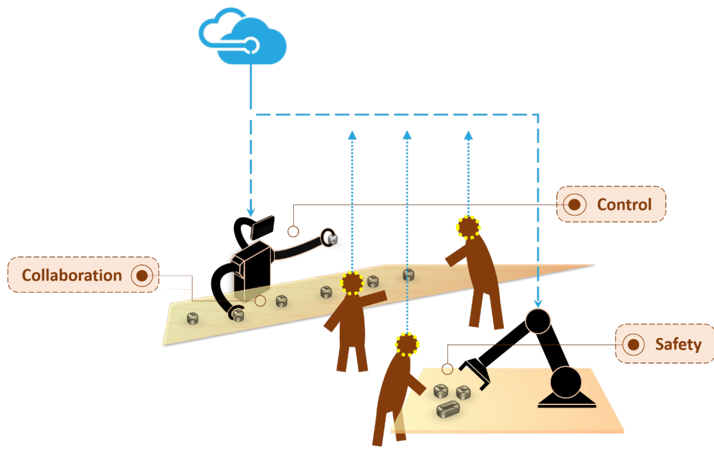


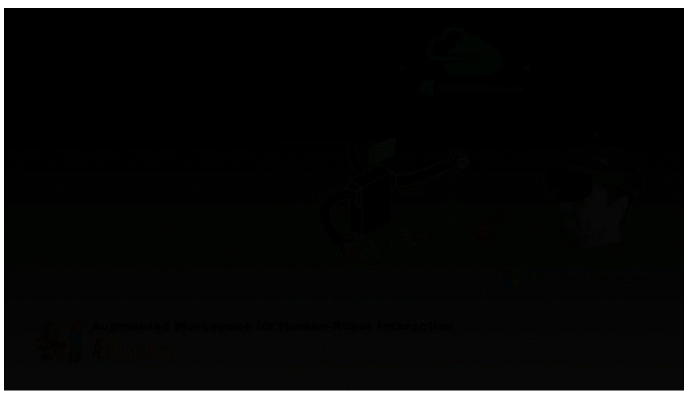






Intention Projection with Hololens





Arizona State University

ASU Home MyASU Ira A. Fulton Schools of

FULLCIRCLE

Faculty Students Alumni Fulton Sch

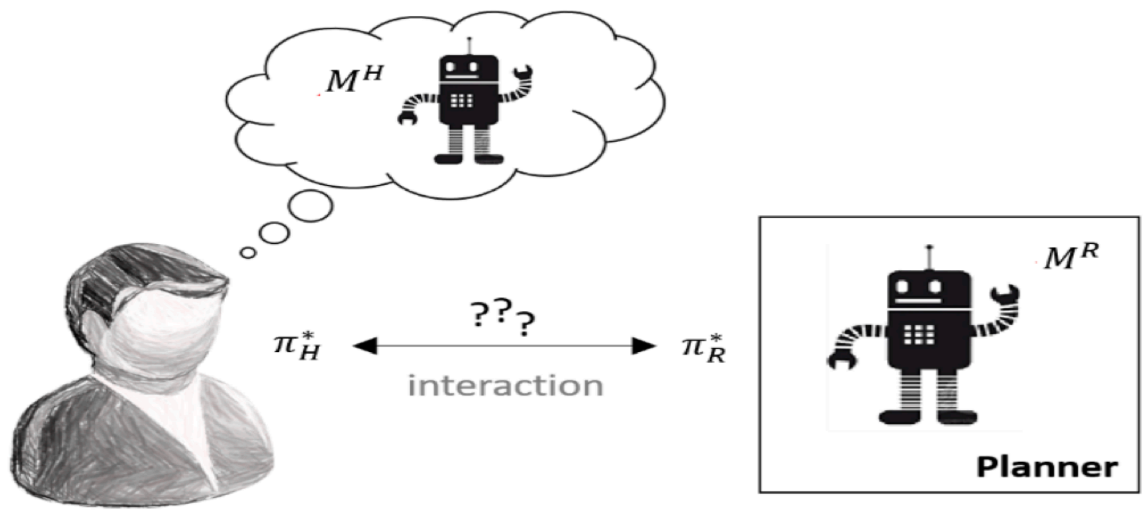
Web Site:
ae-robots.com

ASU TEAM TAKING CONCEPT FOR CLOSER HUMAN-ROBOT CONNECTION TO U.S. IMAGINE CUP FINALS

Posted by Joe Kullman | Apr 6, 2017 | Students

Above: Computer science doctoral students (left to right) Anagha Kulkarni, Sarath Sreedharan and Tathagata Chakraborti have combined aspects of robotics, artificial intelligence, cognitive neuroscience and virtual-reality technology in their project for the Microsoft Imagine Cup competition. Photographer: Marco-Alexis Chaira/ASU.

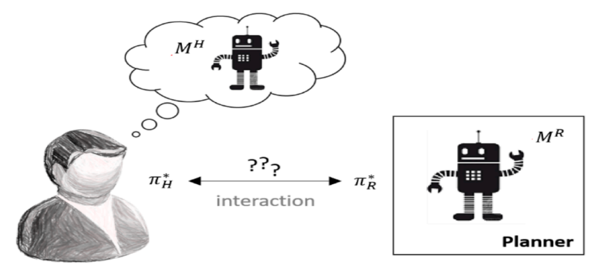
Teaming Requires Modeling the Human's Model of You



Model differences with human in the loop

- The robot and human may have different models of the same task
  - *Consequence →*
    - **Plans that are optimal to the robot may not be so in the model of the human**

      → "Inexplicable" plans

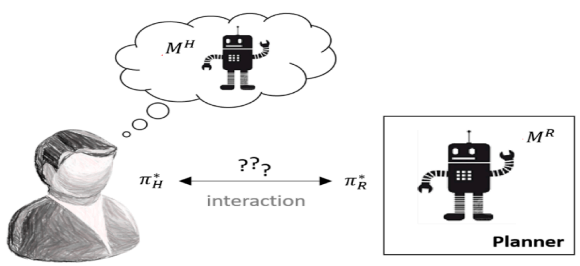


Model differences with human in the loop

- The robot and human may have different models of the same task
  - *Consequence* →
    - **Plans that are optimal to the robot may not be so in the model of the human**

      → ***"Inexplicable" plans***

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



Explicability

A Human-Aware Planning (HAP) Problem is a tuple $\langle \boldsymbol{\mathcal{M}^R}, \boldsymbol{\mathcal{M}^R_h} \rangle$ 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 $\boldsymbol{\mathcal{M}^R_h} = \langle \boldsymbol{D^R_h}, \boldsymbol{I^R_h}, \boldsymbol{G^R_h} \rangle$ 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.

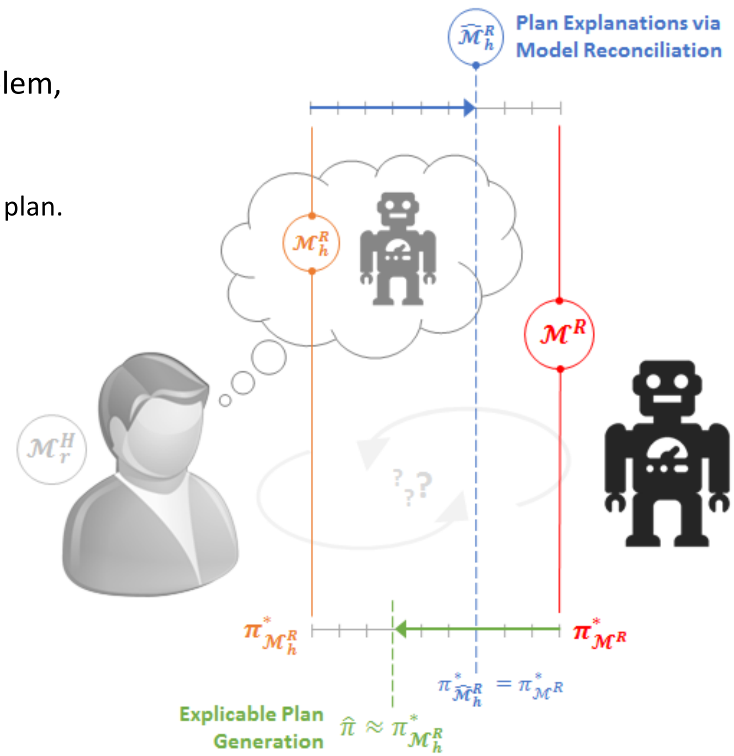
Explicable Plan π →

(1) $\delta_{\mathcal{M}^R}(I^R, \pi) \vDash G^R$

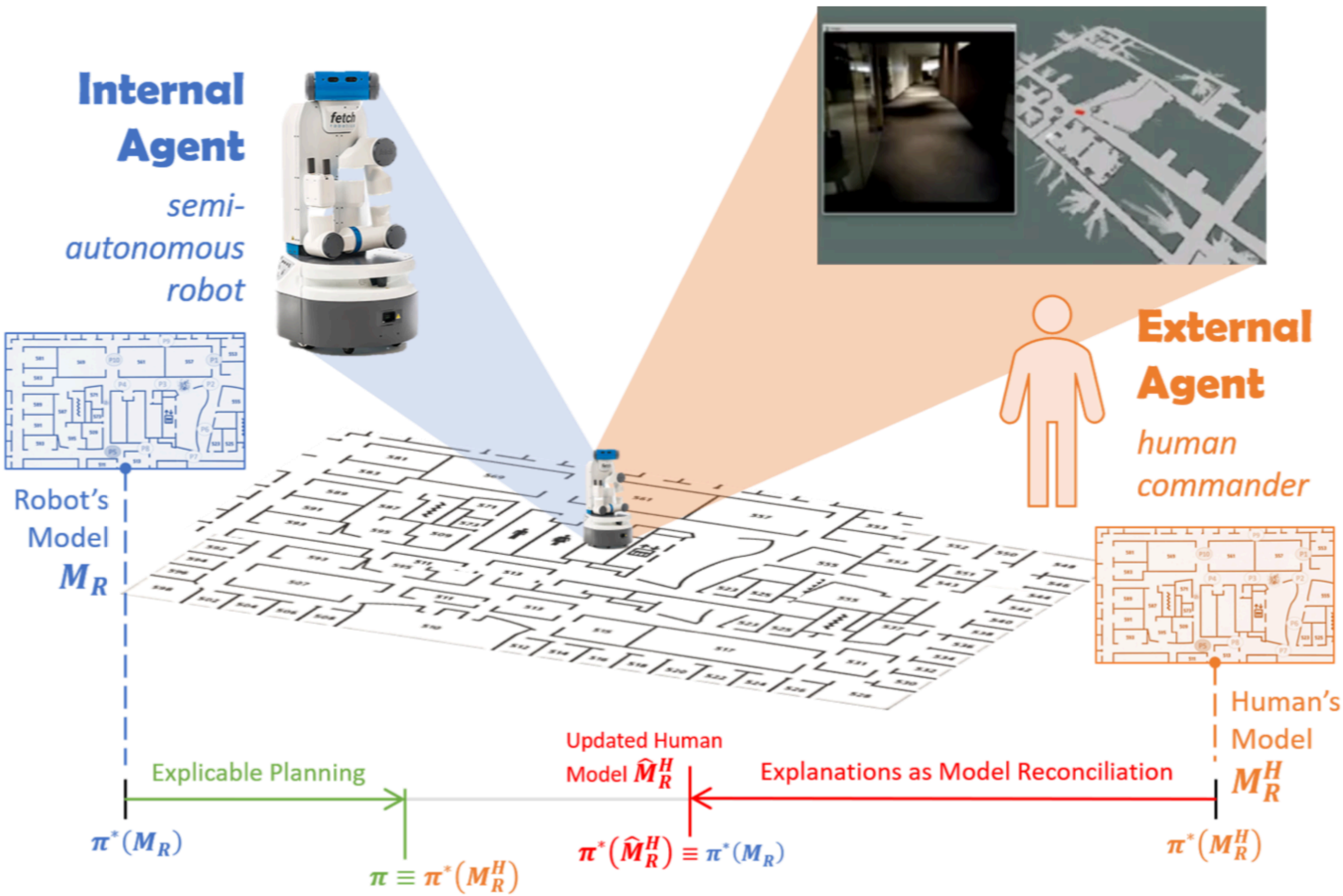
→ is executable in robot's model

(2) $C\left(\pi, \mathcal{M}^R_h\right) \approx C^*_{\mathcal{M}^R_h}$

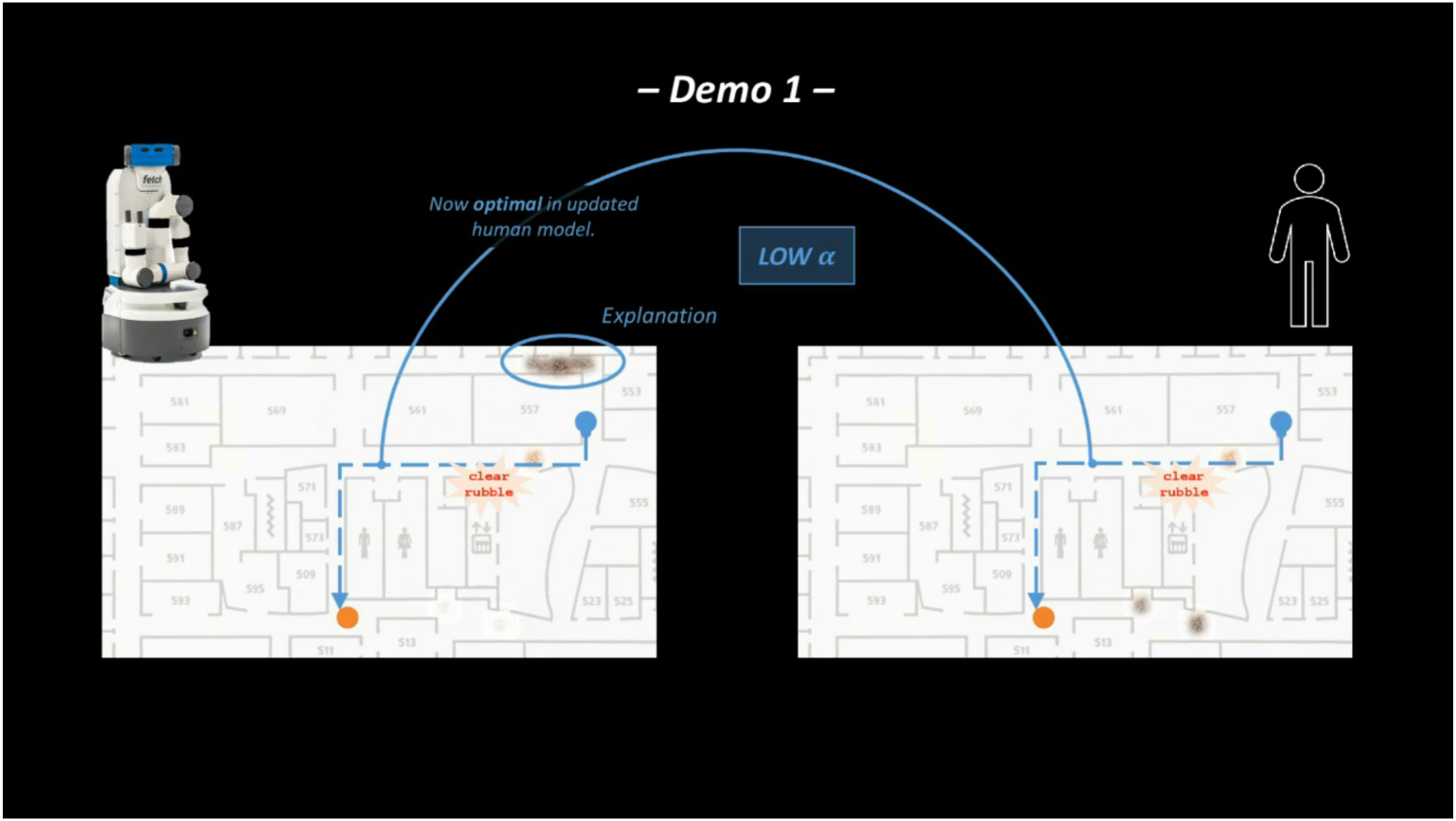
→ *is close to optimal in human's model*



[Plan Explicability for Robot Planning, ICRA 2017]



Internal Agent
semi-autonomous robot
Robot's Model M_R
External Agent
human commander
Human's Model M_R^H
Explicable Planning
Updated Human Model $\widehat{M}_R^H$
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
clear rubble
clear rubble

Explicable Plan

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

$$\underset{\pi_{M_R}}{\operatorname{argmin}} \underbrace{\boxed{cost(\pi_{M_R})}}_{\text{Cost of robot plan}} + \alpha \cdot \underbrace{\boxed{dist(\pi_{M_R}, \pi_{\mathcal{M}_R^*})}}_{\text{Distance between robot plan and human's expectation of robot plan}}$$

Explicable Plan

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

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

Robot does not have access to
human's expectation model

Explicable Plan

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

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

$$\boxed{F} \circ \boxed{\mathcal{L}^*(\pi_{M_R})}$$

F = {$task_1$, $task_2$, $task_3$}

Plan = {a_1, a_2, a_3, ..., a_n}

L^* = $task_1$ No label - Inexplicable $task_2$ $task_1$

Explicable Plan

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

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

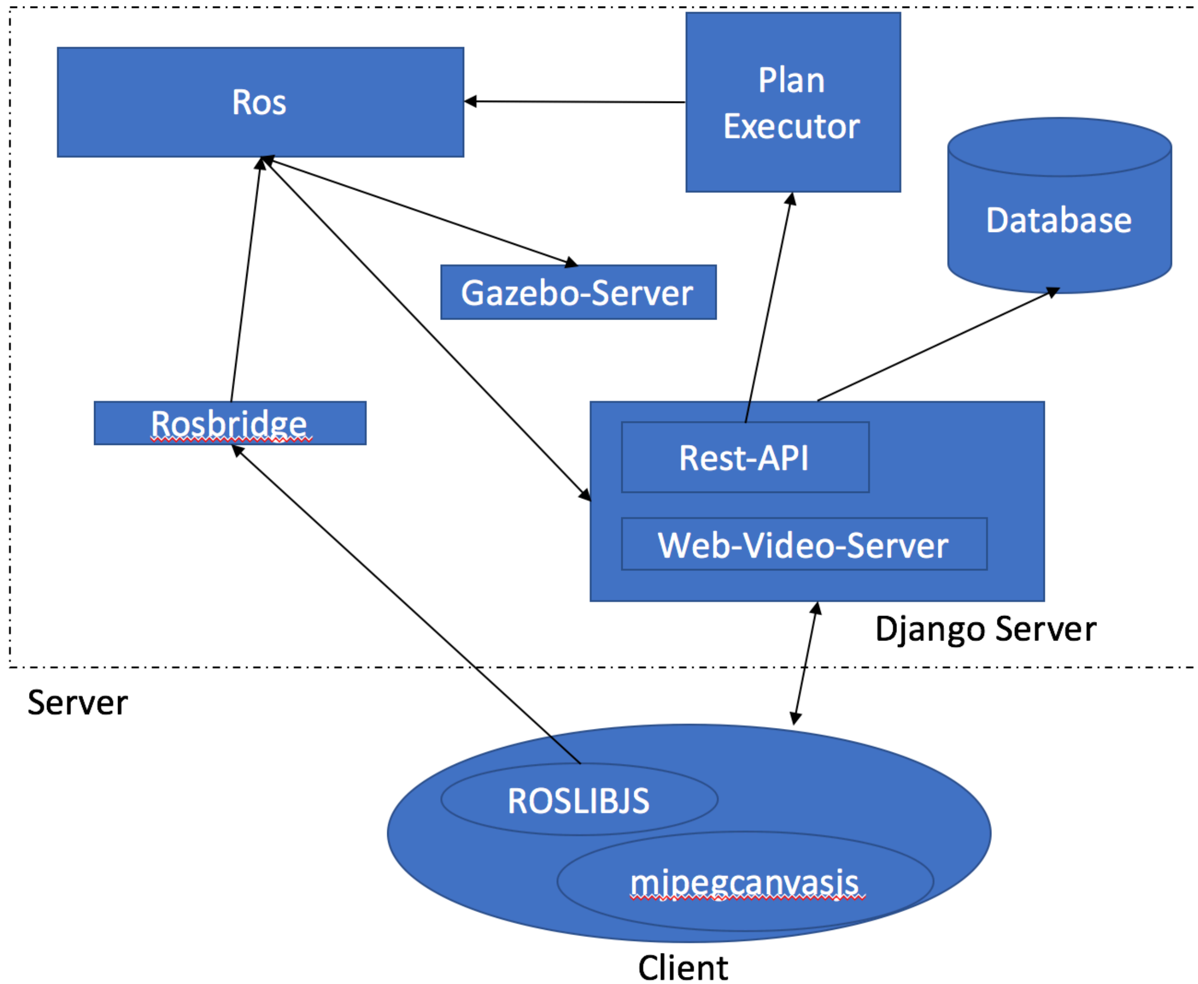
$$\underset{\pi_{M_R}}{\operatorname{argmin}}\, cost(\pi_{M_R}) + \alpha \cdot \boxed{F} \circ \boxed{\mathcal{L}^*_{CRF}(\pi_{M_R} | \{S_i | S_i = \mathcal{L}^*(\pi^i_{M_R})\})}$$

Domain independent function taking plan labels as input

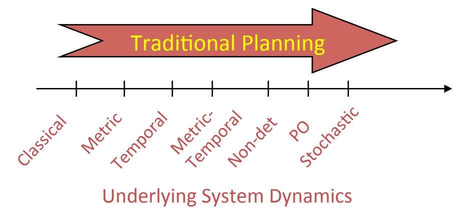
Human's labeling scheme using linear-chain CRFs (Conditional Random Fields)

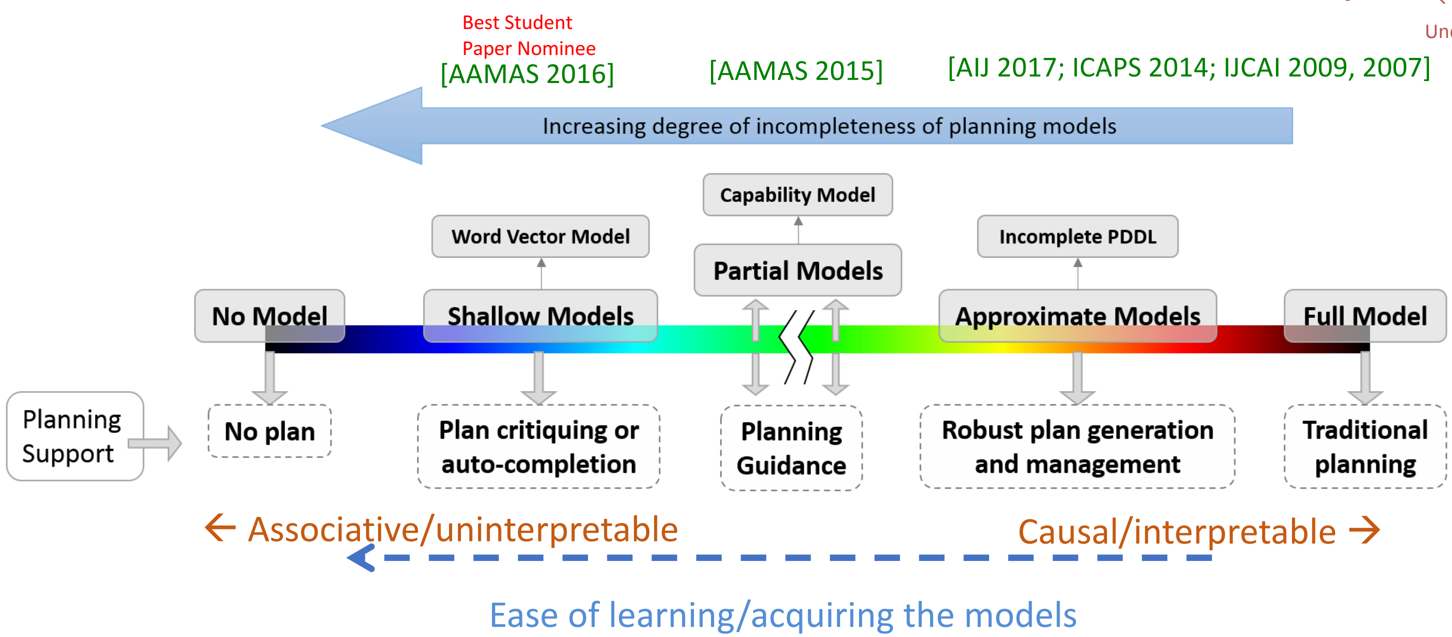
Web Interface to collect human feedback on robot task plans

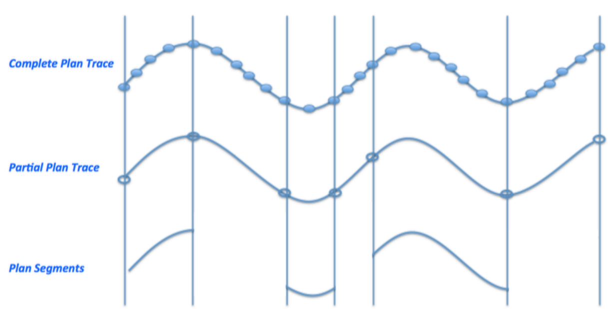
- Goal: Create a web application that enables researchers to leverage crowd sourcing services (eg: mechanical turker) to perform HRI studies in a simulated environment.
- We are specifically interested in enabling users to annotate and/or modify robot task plans being presented to them.
- Related Projects:
  - http://jpdelacroix.com/simiam/
  - http://planit.cs.cornell.edu/
  - http://robotwebtools.org/



Learning: A Spectrum of Domain Models







UNCLASSIFIED

Note the contrast to ML research where the progress is going from uninterpretable/non-causal models *towards* interpretable and causal models.
So we might meet in the middle!

Action Vector Models can be used to Recognize Plans

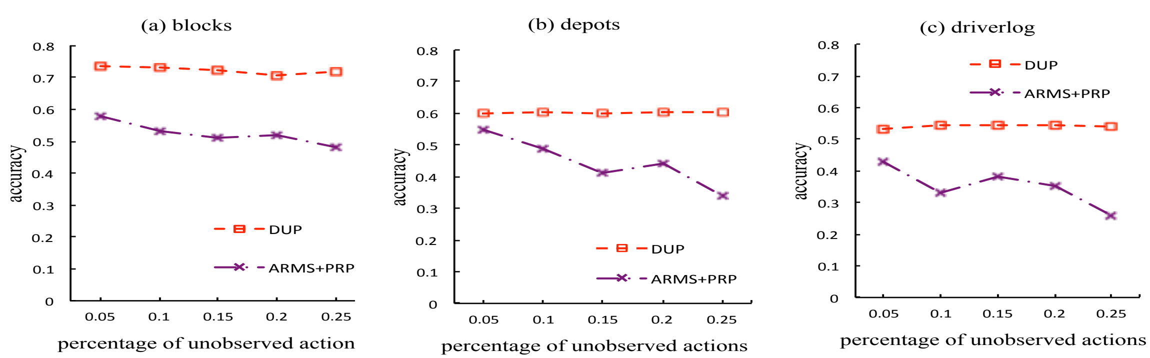
With the learnt vectors w_i, we can predict the target plan (as the most consistent with the affinities). We use an EM procedure to speedup the prediction.

$$\mathcal{F}(\tilde{p}) = \sum_{k=1}^{M} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{k+j} | w_k)$$

- M = |the target plan|

The target plan to be recognized

Learning shallow models can avoid overfitting!!



Algorithm 1 Framework of our DUP algorithm

Input: plan library $\mathcal{L}$, observed actions $\mathcal{O}$
Output: plan $\tilde{p}$

learn vector representation of actions
initialize $\Gamma_{o,k}$ with $1/M$ for all $o \in \bar{\mathcal{A}}$, when k is an unobserved action index
while the maximal number of repetitions is not reached **do**
  sample unobserved actions in $\mathcal{O}$ based on Γ
  update Γ based on Equation (6)
  project Γ to [0,1]
end while
select actions for unobserved actions with the largest weights in Γ
return $\tilde{p}$

Nominated for Best Student Paper Award at [AAMAS16]

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

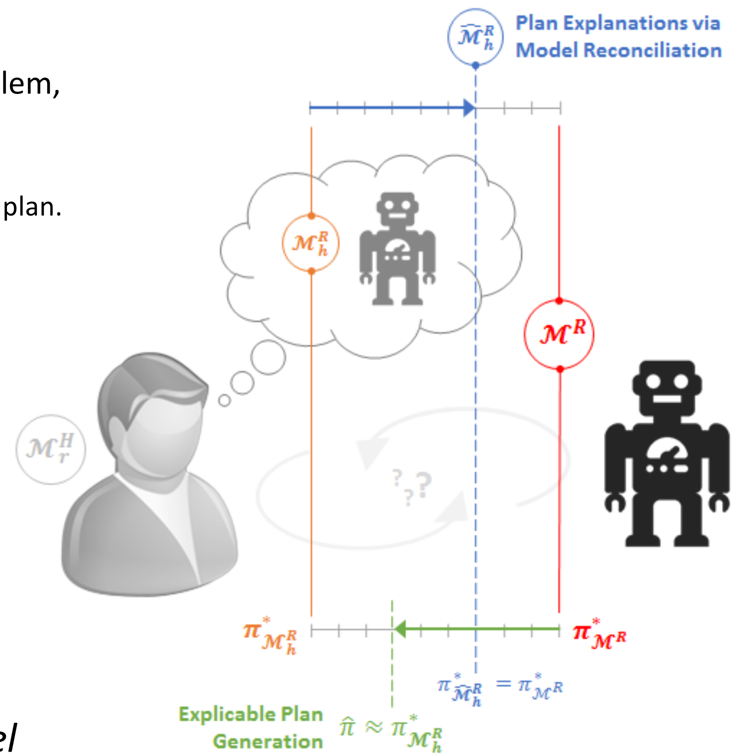
→ is a model update to the human

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

→ π is optimal in robot's model

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

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

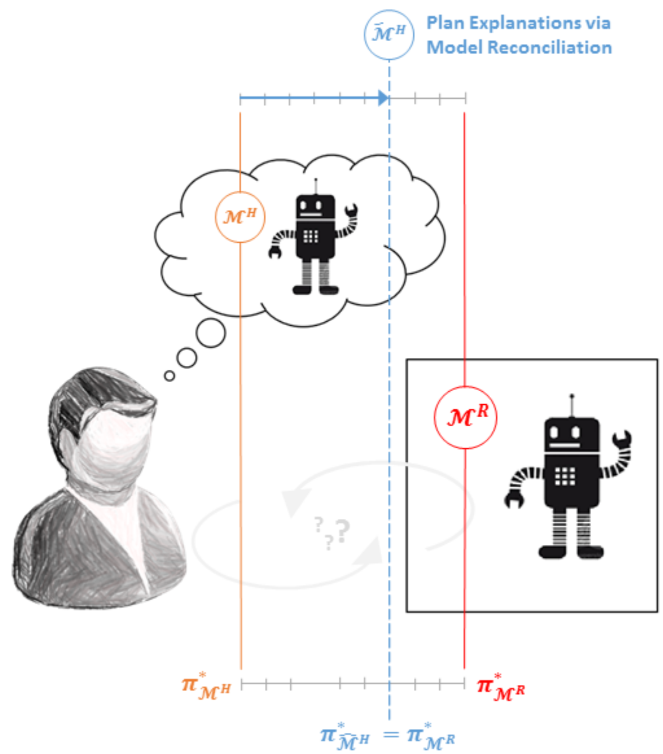


[Moving Beyond Explanation as Soliloquy; IJCAI 2017]

Explanations as Model Reconciliation

- "XAI" is hot.. But mostly as a debugging tool for "inscrutable" representations
  - "Pointing" explanations
- Explanations are critical for collaboration .. But they are not a *soliloquy* by the agent
- Model Reconciliation view hews close to psychological theories, e.g. [Lombrozo, 2006]

- Constraints for reasoning
- Contrastive property
- Soundness and Completeness
- Account human model



[Moving Beyond Explanation as Soliloquy; IJCAI 2017]

Different Kinds of Explanations

- **Model Patch Explanation (MPE)**
  - All the model differences.
- **Plan Patch Explanation (PPE)**
  - Model differences pertaining to actions in the plan (plan is at least executable after this).
- **Minimally Complete Explanation (MCE)**
  - Minimum number of corrections to the human model that makes the given plan optimal in the update model.
- **Minimally Monotonic Explanation (MME)**
  - Minimum number of updates to human model so that plan remains optimal irrespective of future problems.
- ***Approximate* Minimally Complete Explanations**
  - Approximate solution to MCE using only necessary condition for optimality of given plan in updated model.

Example - FetchWorld

Robot Model

```
(: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)))
```

PPE = MPE

MCE

MME

Human Model of Robot

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

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

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

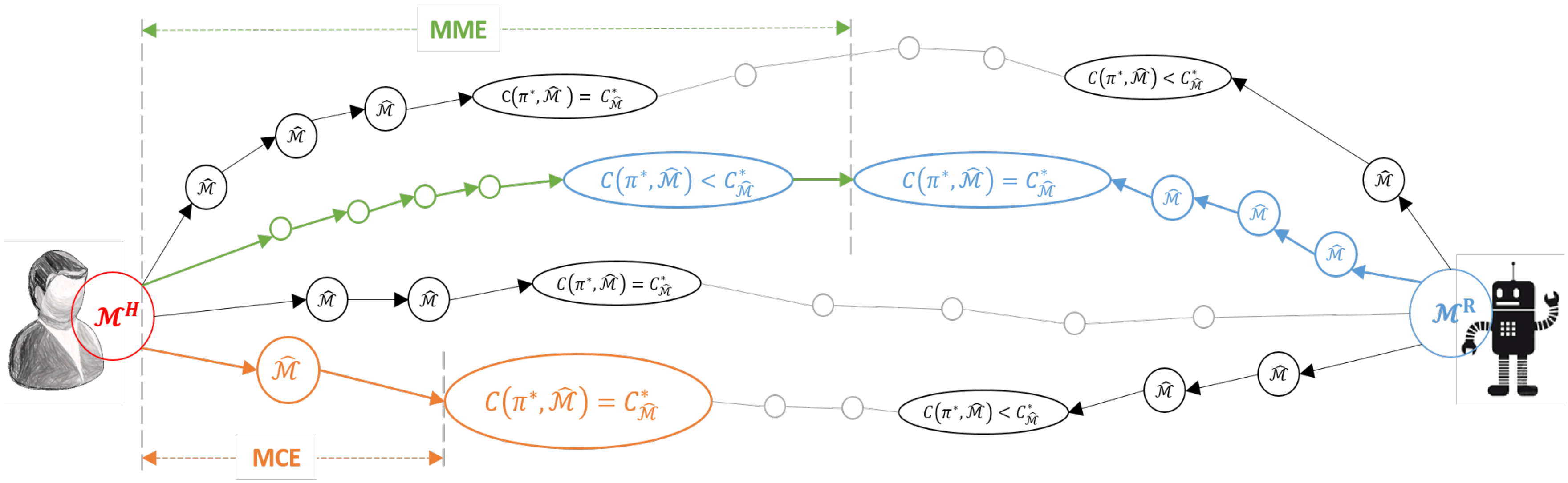
```
Problem:
(:init (block-at b1 loc1) (robot-at loc1) (hand-empty))
(:goal (and (block-at b1 loc2)))

Robot's Optimal Plan:
pick-up b1 -> tuck -> move loc1 loc2 -> put-down b1
              ??

Human's Expected Plan:
pick-up b1 -> move loc1 loc2 -> put-down b1
```

Different Kinds of Explanations

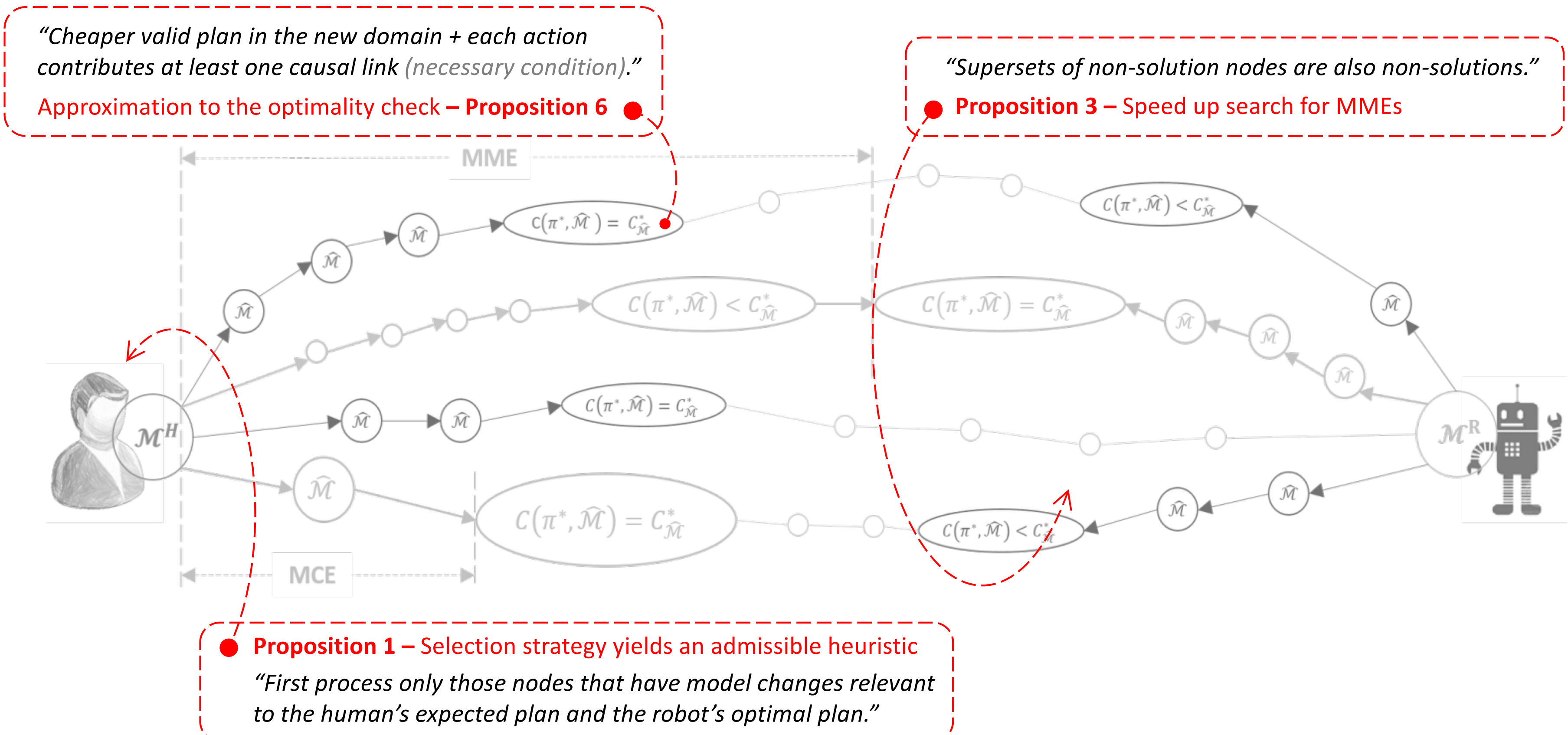
Explanation Type	Completeness	Conciseness	Monotonicity	Computability
Model Patch Explanation (MPE)	✓	✗	✓	✓
Plan Patch Explanation (PPE)	✗	✓	✗	✓
Minimally Complete Explanation (MCE)	✓	✓	✗	?
Minimally Monotonic Explanation (MME)	✓	✓	✓	?
Approximate Minimally Complete Explanations	✗	✓	✗	✓

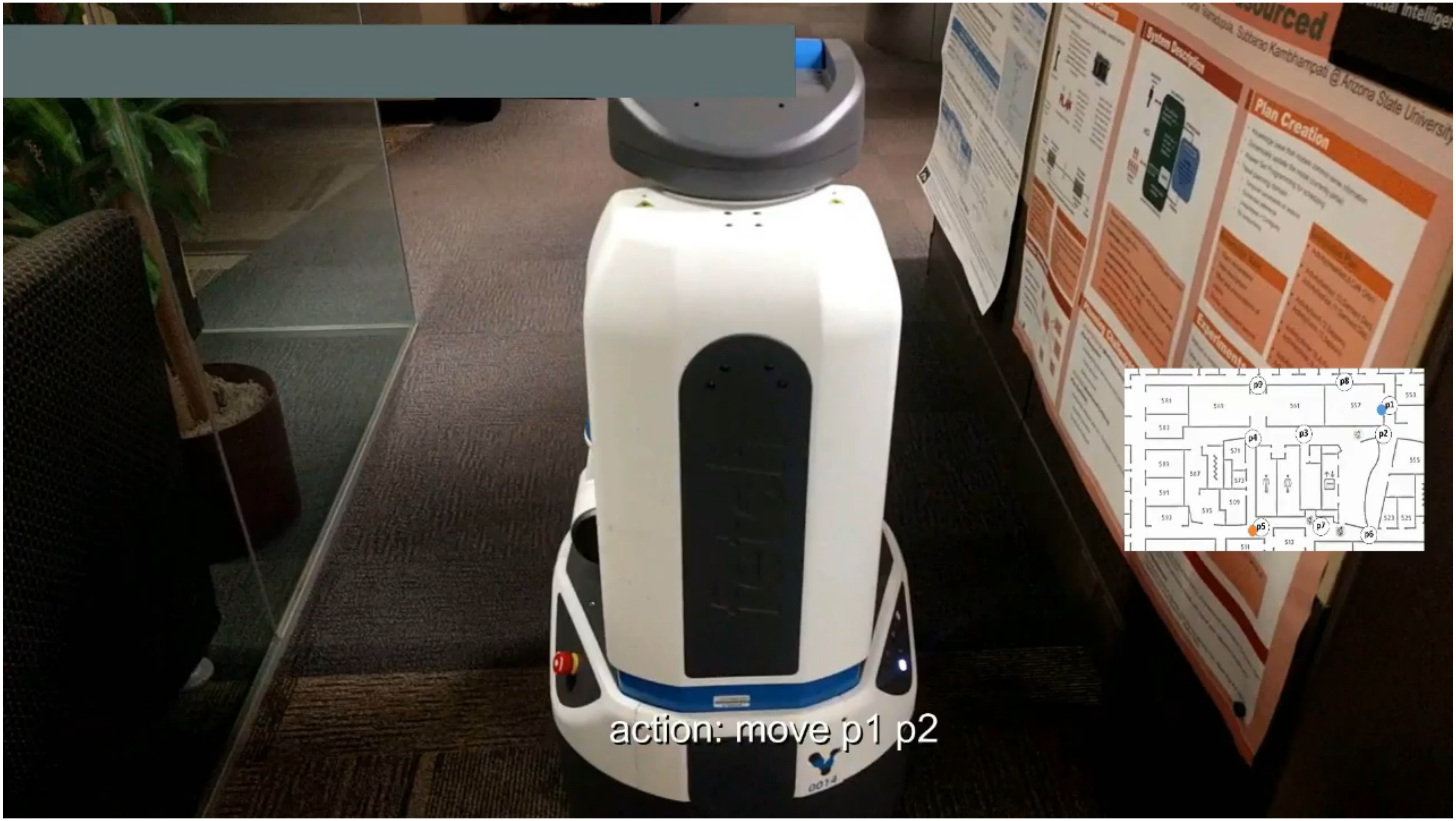
- Note that these requirements are often at odds with each other - an explanation that is very easy to compute may be very hard to comprehend.
- We minimize the size (and increase the comprehensibility) of explanations by not exposing information that is not relevant to the plan being explained while still satisfying as many requirements as possible.

Model Space Search for Model Reconciliation



Model Space Search for Model Reconciliation




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

Trading Explicability & Explanation

- What does this mean for planning?
  - **The robot (planner) has to decide in which model it is planning in.**
    - Trade-off cost of explaining versus cost of suboptimality → *model space search*

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

→ ϵ is a model update to the human

(2) $\delta_{\mathcal{M}^R}(I^R, \pi) \models G^R$

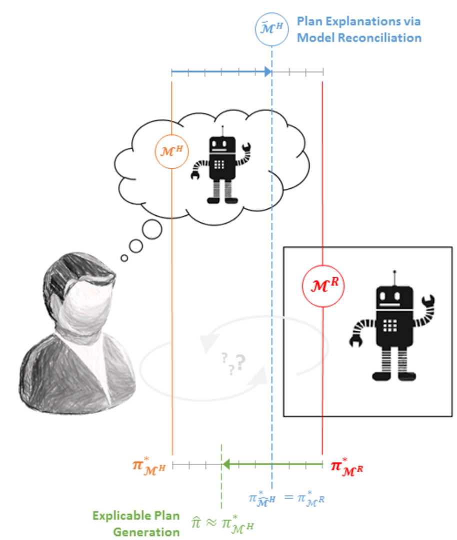
→ π is executable in robot's model

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

→ π is optimal in the updated human model

(4) $\pi = argmin_\pi \left\{|\epsilon| + \alpha \times \left|C(\pi, \mathcal{M}^R) - C^*_{\mathcal{M}^R}\right|\right\}$

→ trade-off costs of explanation versus explicability

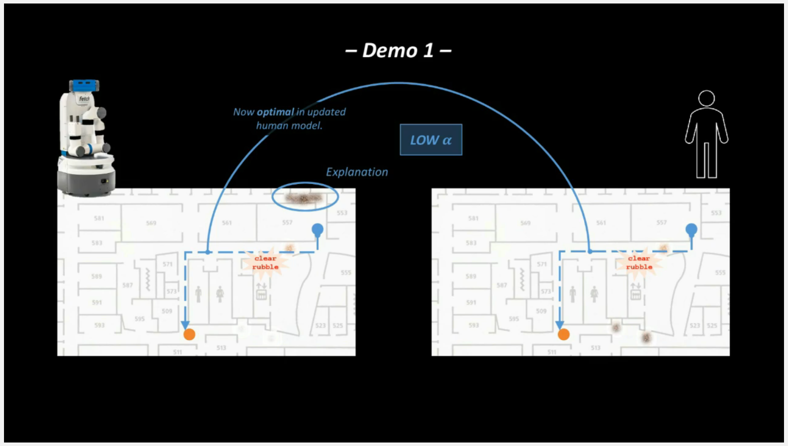


[AAAI Fall Symposium, 2017]

Explicability/Explanation Tradeoff in Action

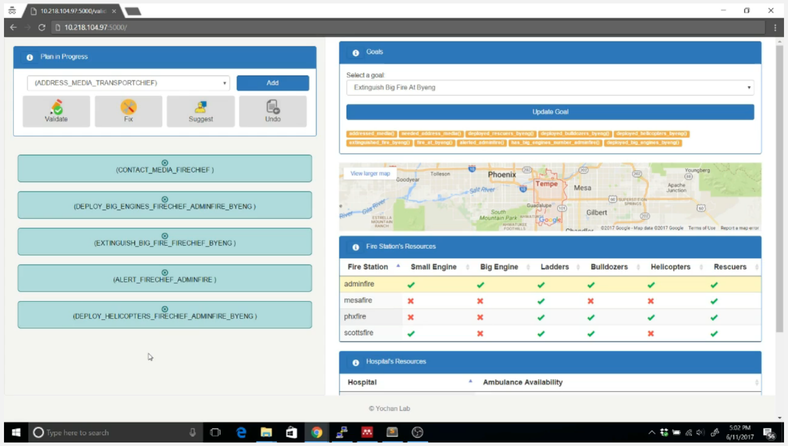
Search & Reconnaissance scenario with an internal semi-autonomous agent and an external human supervisor.

- Combines explanations + explicability.
- *To be presented at AAAI 2017 Fall Symposium on AI-HRI*



Decision Support scenario with *human planners* who are making disaster response strategies in the control room.

- Iterative reconciliation of models.
- *Appeared in ICAPS'16 System Demos.*



↑ **Situational Awareness**
↓ **Information Overload**

[AAAI Fall Symp, 2017]

Are we in the right direction?

- Let's ask Humans

- (It is hard for AI to say we are pro-human, if we are oblivious to IRB..)
- (IRB guidelines themselves may have to evolve with advances in Human-aware AI)

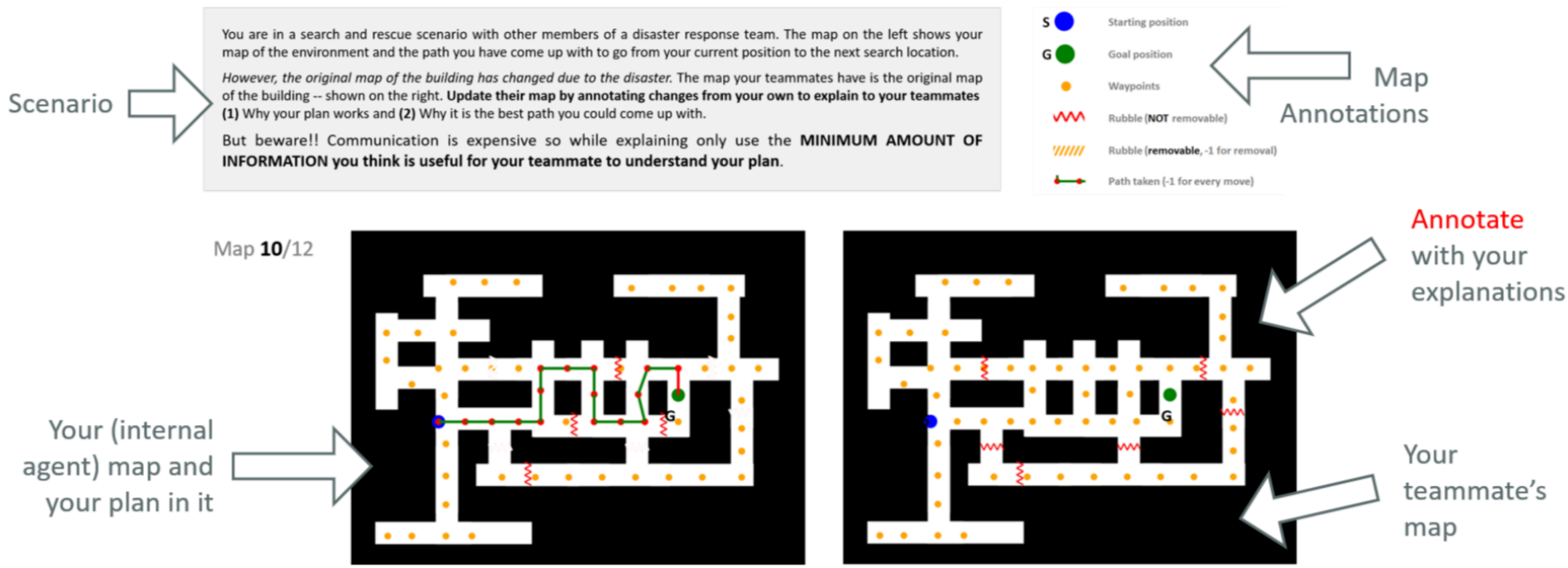
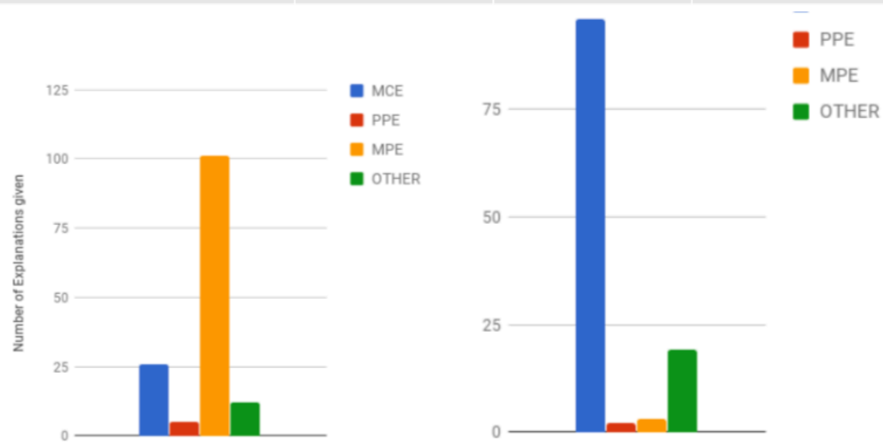


Figure 3: Interface for Study-1 where participants assumed the role of the internal agent and were asked to explain their plan to a teammate with a possibly different model or map of the world.

Different Kinds of Explanations

Explanation Type	Completeness	Conciseness	Monotonicity	Computability
Model Patch Explanation (MPE)	✓	✗	✓	✓
Plan Patch Explanation (PPE)	✗	✓	✗	✓
Minimally Complete Explanation (MCE)	✓	✓	✗	?
Minimally Monotonic Explanation (MME)	✓	✓	✓	?
Approximate Minimally Complete Explanations	✗	✓	✗	✓



(a) Study-1:C1 (b) Study-1:C2

Figure 4: Count of different types of explanations for Study-1 conditions C1 and C2.

Figure 7: Interface for Study-2 where participants assumed the role of the external commander and evaluated plans provided by the internal robot. They could request for plans and explanations to those plans (e.g. if not satisfied with it) and rate those plans as optimal or suboptimal based on that explanation. If still unsatisfied with the plan even after the explanation they could chose to pass and move on to the next problem.

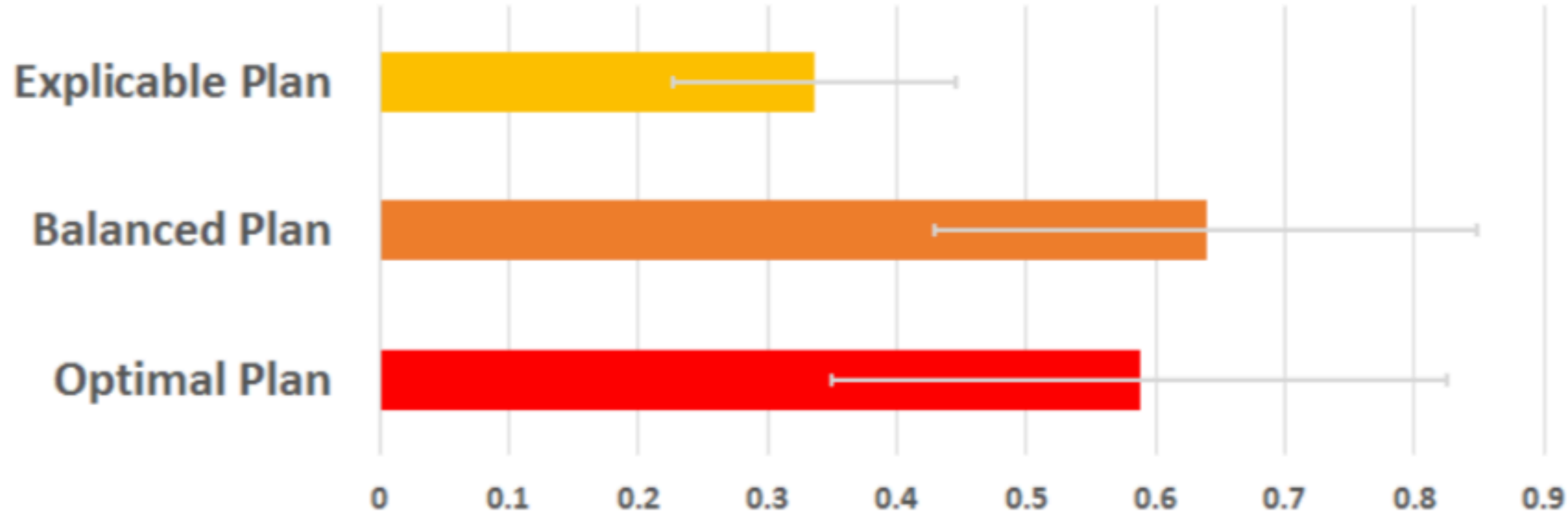
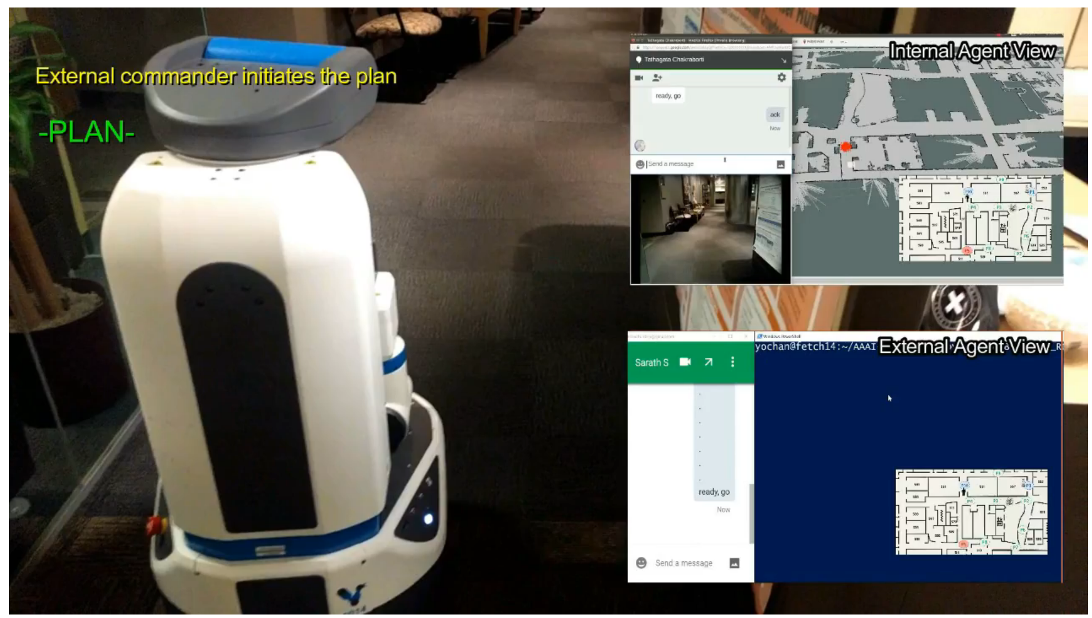


Figure 10: Percentage of times explanations were sought for in Study–2 when participants presented with explicable plans vs. balanced or robot optimal plans with explanations.

Handling Multiple Humans & Differing Abstractions

- Handling Multiple Human Agents (or single agent with incomplete model)
    - An interesting mapping to "Conformant Planning" setting

- Handling models that are at different levels of abstraction
    - E.g. A doctor "explains" her diagnosis to a colleague in a different way than to a patient.



[ICAPS 2018]

Summary our research

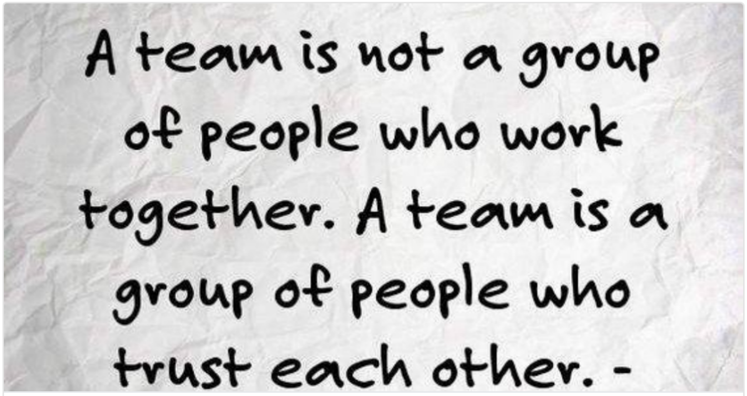
- Effective human-robot teaming requires that the robot model the human's goals and intentions <u>as well as the human's model of robot's capabilities</u>
- Such a model is needed to show *explicable behavior* (i.e., behavior that the human expects from the robot), to the extent possible
- And provide *explanations* when explicability is not possible
  - Explanations cannot be *soliloquy*
  - They are best modeled as "model reconciliation"
- It is possible to tradeoff explicability and explanation
- ..and to model multiple humans or differing abstraction levels

Objective of this talk..

- Why isn't human-aware AI all over the place already?
- Why we should pursue it? (Hint: It broadens the scope & promise of AI)
- Research Challenges in HAAI (Case Study: Our research on Human-aware Planning & Decision Making)
- Long term issues (Trust); Ethical Dilemmas

Implications for "Trust in Autonomy"

- One holy-grail in human aware AI systems is engendering trust in the humans
- The mechanisms of long term trust are complex
- However, ability of the agent to show ***explicable behavior*** and provide comprehensible ***explanations*** are clearly critical for engendering trust
- (Other factors: Assessment of self-competence and human competence)



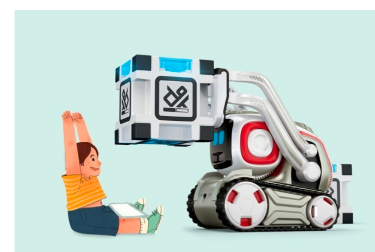
"As soon as a tool becomes a partner, thousands of years of evolutionary conditioning is brought to bear on our interactions with it.."

--Daniel Fessler (UCLA Anthropologist)

(New) Ethical Quandaries of HAAI

Every tool is a weapon, if you hold it right..
--Ani Difranco

- Evolutionarily, mental modeling allowed us to both cooperate or compete/sabotage each other
  - Lying is possible only because we can model others' mental states!
- HAAI systems with mental modeling capabilities bring additional ethical quandaries
  - E.g. Automated negotiating agents that misrepresent their intentions to gain material advantage
  - Your personal assistant that tells you white lies to get you eat healthy (…or not..)
- Humans' example closure tendencies are more pronounced for emotional/social intelligence aspects
  - No one who saw Shakey the first time thought it could shoot hoops; yet the first people interacting with Eliza assumed it is a real doctor!
  - Concerns about HAAI "toys" such as Cozmo (e.g. Sherry Turkle)



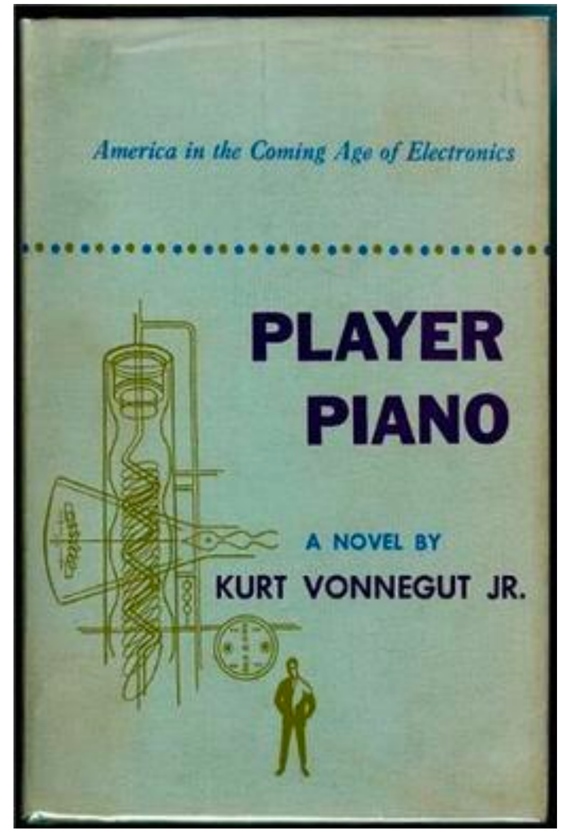
[On Mental Modeling & Acceptable Symbiosis in Human-AI Collaboration; arXiv 1801.09854]

HAAI Brings in a slew of additional challenges

"If only it weren't for the people, the goddamned people," said Finnerty,
"always getting tangled up in the machinery.
If it weren't for them, earth would be an engineer's paradise."

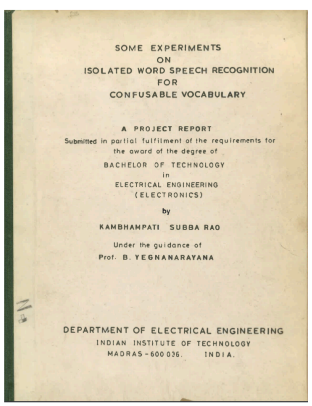
--From Player Piano by Kurt Vonnegut, Jr.

..but perhaps they are worth our time to tackle.
after all, some of our best friends are human…



The Fundamental Questions Facing Our Age

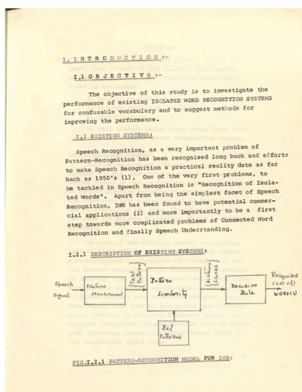
- Origin of the Universe
- Origin of Life
- Nature of Intelligence



SOME EXPERIMENTS
ON
ISOLATED WORD SPEECH RECOGNITION
FOR
CONFUSABLE VOCABULARY

A PROJECT REPORT
Submitted in partial fulfilment of the requirements for
the award of the degree of
BACHELOR OF TECHNOLOGY
in
ELECTRICAL ENGINEERING
(ELECTRONICS)
by
KAMBHAMPATI SUBBA RAO
Under the guidance of
Prof. B. YEGNANARAYANA

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
MADRAS - 600 036. INDIA.



I. INTRODUCTION:-

I.1 OBJECTIVE:-

The objective of this study is to investigate the performance of existing ISOLATED WORD RECOGNITION SYSTEMS for confusable vocabulary and to suggest methods for improving the performance.

I.2 EXISTING SYSTEMS:

Speech Recognition, as a very important problem of Pattern-Recognition has been recognised long back and efforts to make Speech Recognition a practical reality date as far back as 1950's (1). One of the very first problems, to be tackled in Speech Recognition is "Recognition of Isolated Words". Apart from being the simplest facet of Speech Recognition, IWR has been found to have potential commercial applications (2) and more importantly to be a first step towards more complicated problems of Connected Word Recognition and finally Speech Understanding.

I.2.1 DESCRIPTION OF EXISTING SYSTEMS:

FIG.I.2.1 PATTERN-RECOGNITION MODEL FOR IWR:

1983 Bachelors thesis ☺

..and the end of all our exploring will be to
arrive where we started
and know the place for the first time.
T.S. Eliot

Summary of the talk..

- Why isn't human-aware AI all over the place already?
- Why we should pursue it? (Hint: It broadens the scope & promise of AI)
- Research Challenges in HAAI (Case Study: Our research on Human-aware Planning & Decision Making)
- Long term issues (Trust); Ethical Dilemmas

Summary our research

- Effective human-robot teaming requires that the robot model the human's goals and intentions as well as the human's model of robot's capabilities
- Such a model is needed to show *explicable behavior* (i.e., behavior that the human expects from the robot), to the extent possible
- And provide *explanations* when explicability is not possible
  - Explanations cannot be *soliloquy*
  - They are best modeled as "model reconciliation"
- It is possible to tradeoff explicability and explanation
- ..and to model multiple humans or differing abstraction levels

[All relevant papers available @ rakaposhi.eas.asu.edu/papers.html]