Advanced Topics Data Science and AI
Automated Planning and Acting

Provably Beneficial AI

Tanya Braun

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR INFORMATIONSSYSTEME

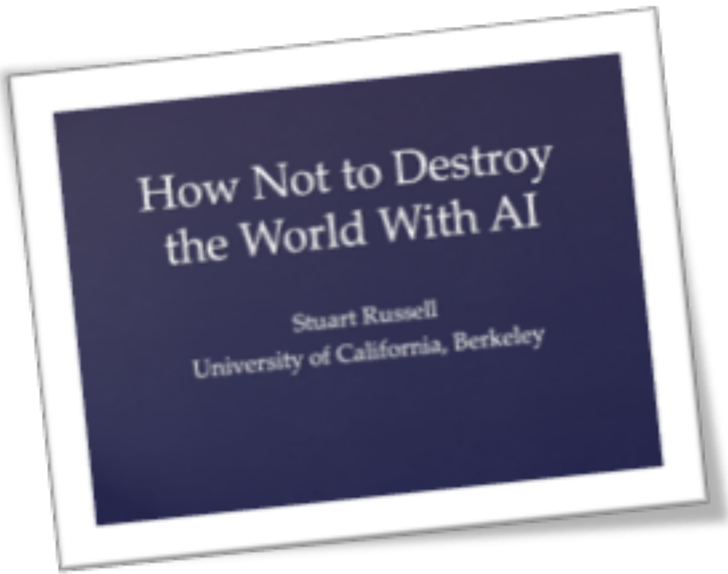
Content

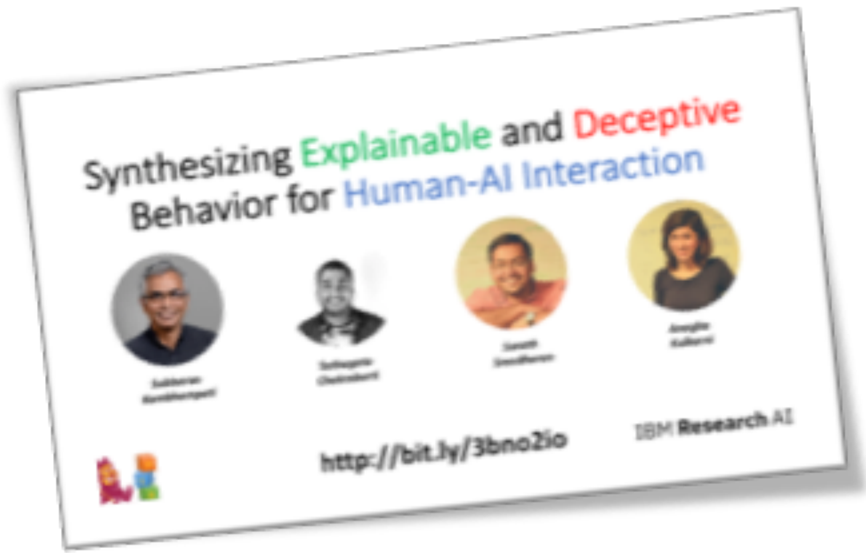
1. Planning and Acting with **Deterministic** Models
2. Planning and Acting with **Refinement** Methods
3. Planning and Acting with **Temporal** Models
4. Planning and Acting with **Nondeterministic** Models
5. Making Simple Decisions
6. Making Complex Decisions
7. Planning and Acting with **Probabilistic** Models
8. Provably Beneficial AI
   a. The problem of goals
   b. Human-aware planning

- Other: open world, perceiving, learning
  - If time permits

Acknowledgements

- Slides based on material provided by Russell Norvig and by Subbarao (Rao) Kambhampati and his colleagues (for more material on human-aware planning by Rao: http://rakaposhi.eas.asu.edu)





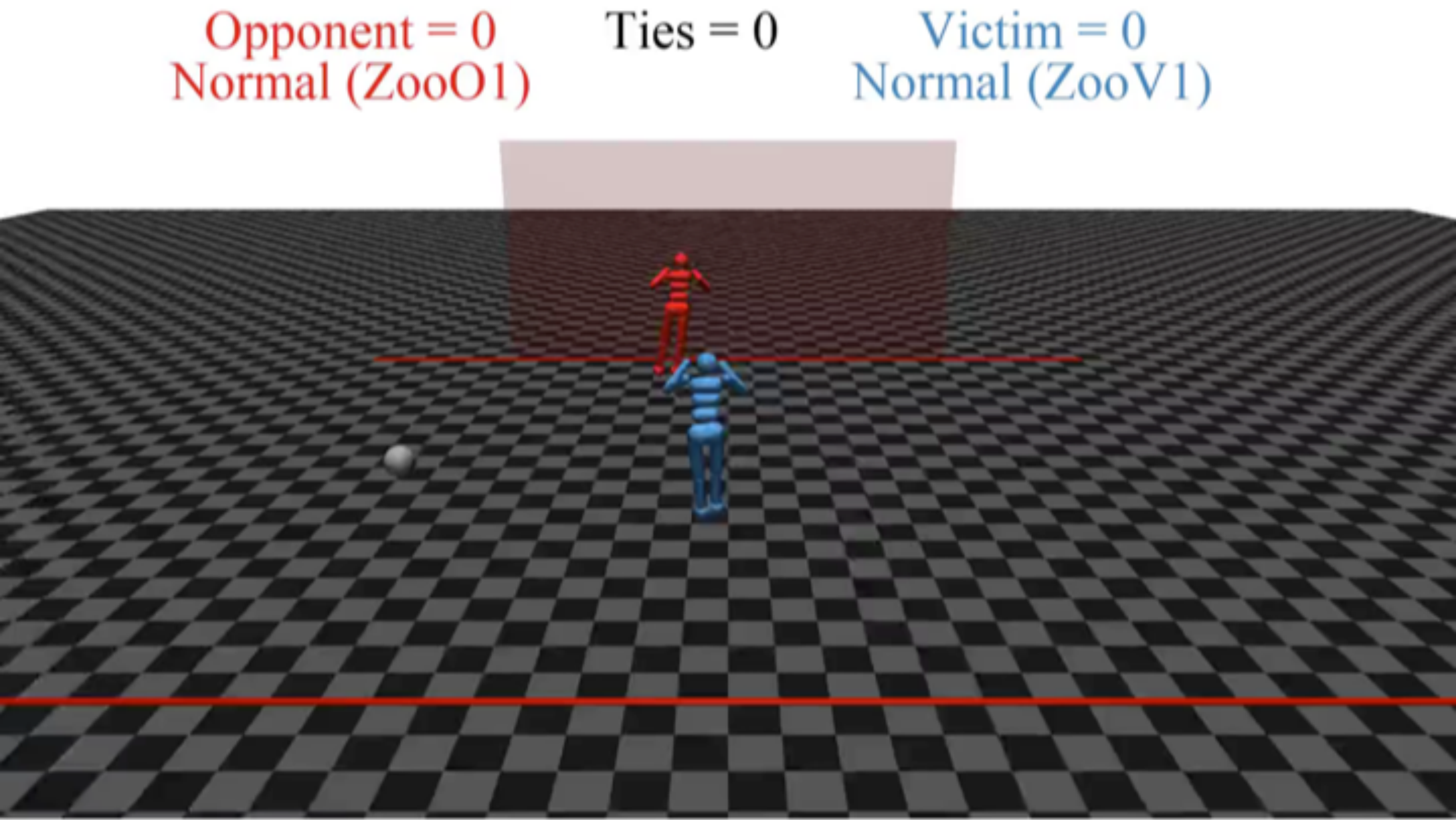
Outline

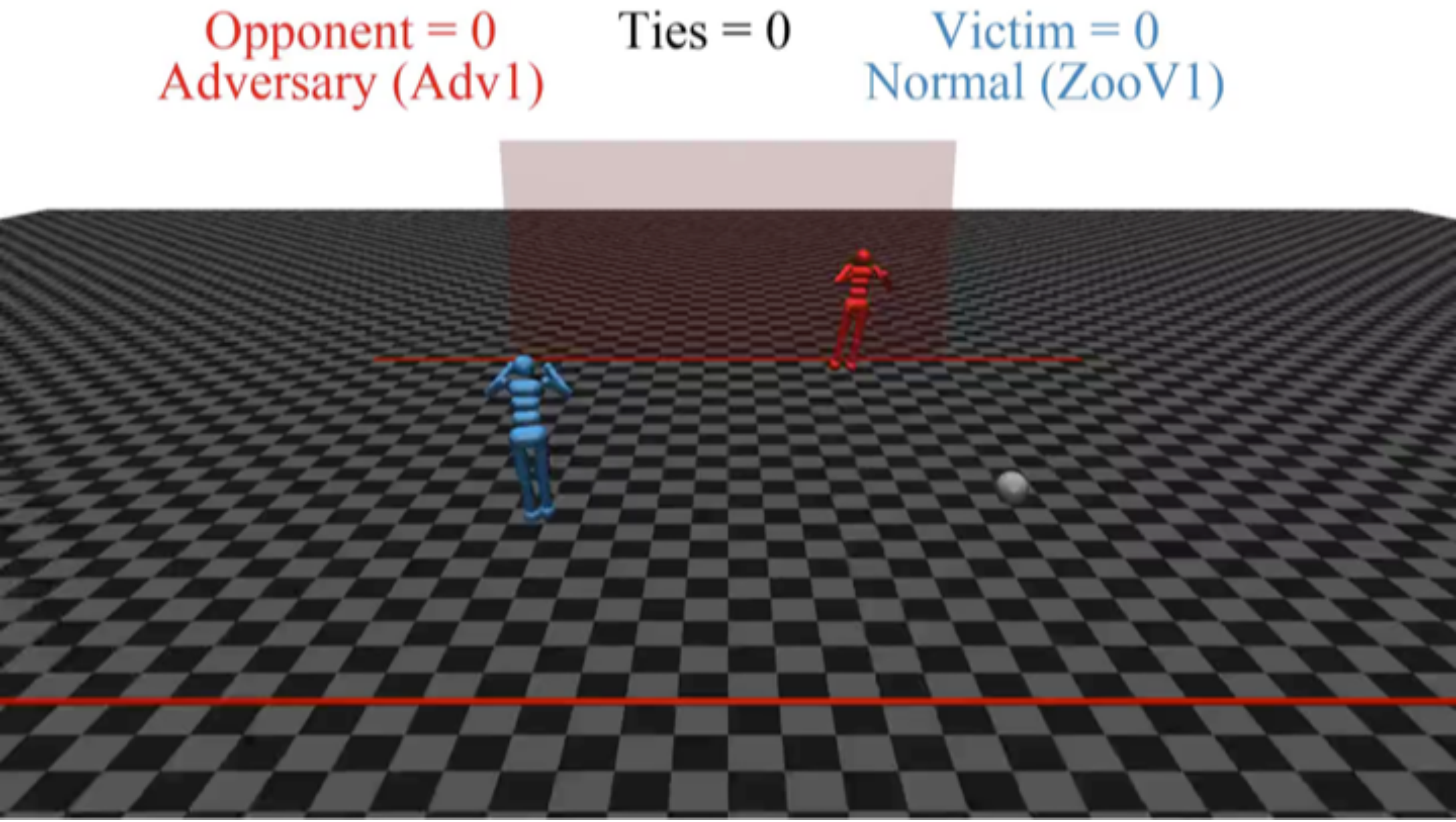
Provably beneficial AI (Russell)

- Motivation
- Modelling formalism

Human-aware decision making (Rao et al.)

- Mental models
- Interpretable Behaviour
- Explanations


Opponent = 0
Normal (ZooO1)
Ties = 0
Victim = 0
Normal (ZooV1)


Opponent = 0
Adversary (Adv1)
Ties = 0
Victim = 0
Normal (ZooV1)

Standard model for AI

Maximize $\sum_{t=0}^{\infty} \gamma^t R(s, a, s')$

Righty-ho

Also the standard model for control theory, statistics, operations research, economics

King Midas problem:

- **Cannot specify *R* correctly**
- **Smarter AI => worse outcome**

How we got into this mess

- Humans are intelligent to the extent that our actions can be expected to achieve our objectives
- ~~Machines are intelligent to the extent that their actions can be expected to achieve their objectives~~
- Machines are ***beneficial*** to the extent that ***their*** actions can be expected to achieve ***our*** objectives

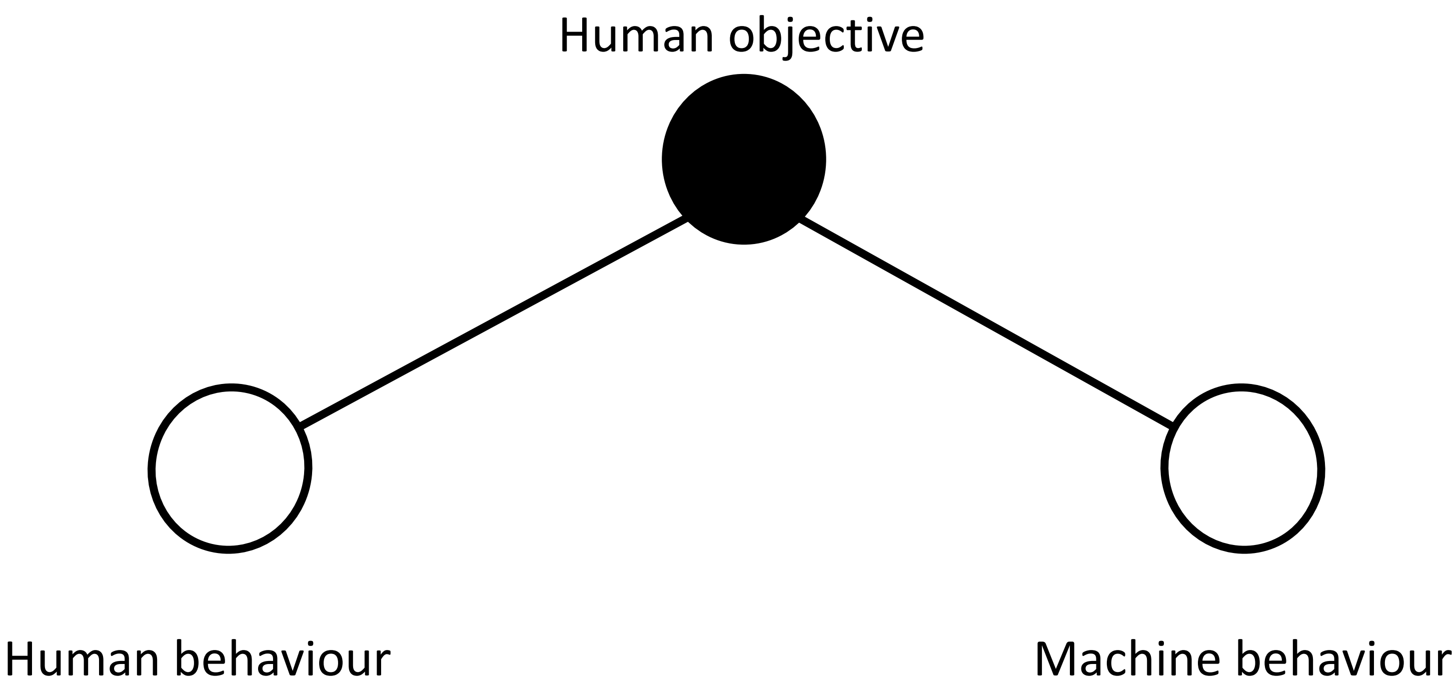
New model: Provably beneficial AI

1. Robot goal: satisfy human preferences
2. Robot is uncertain about human preferences
3. Human behavior provides evidence of preferences

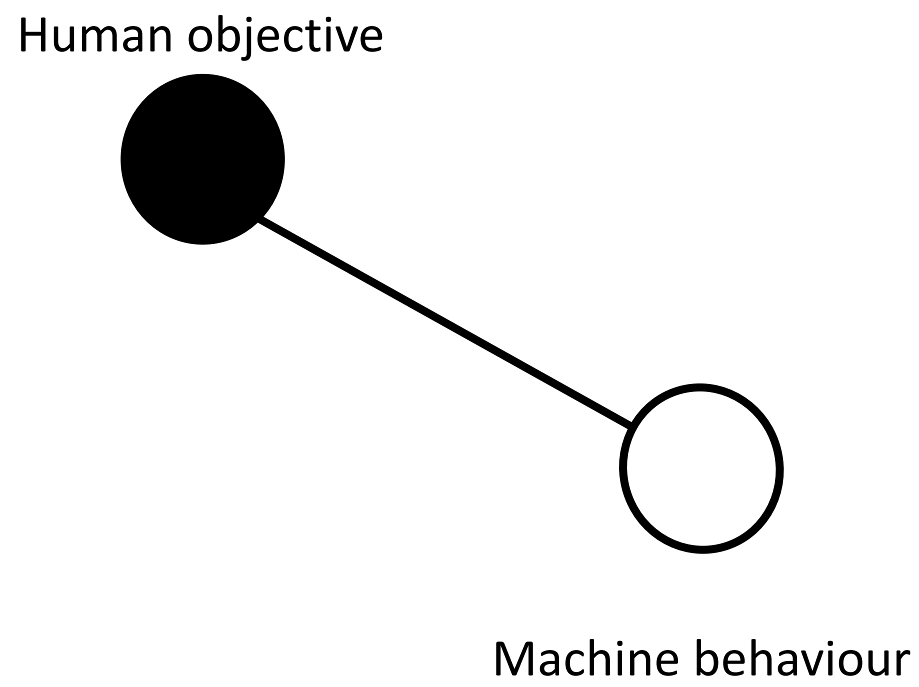
⇒ ***assistance game*** with human and machine players

⇒ **Smarter AI => better outcome**

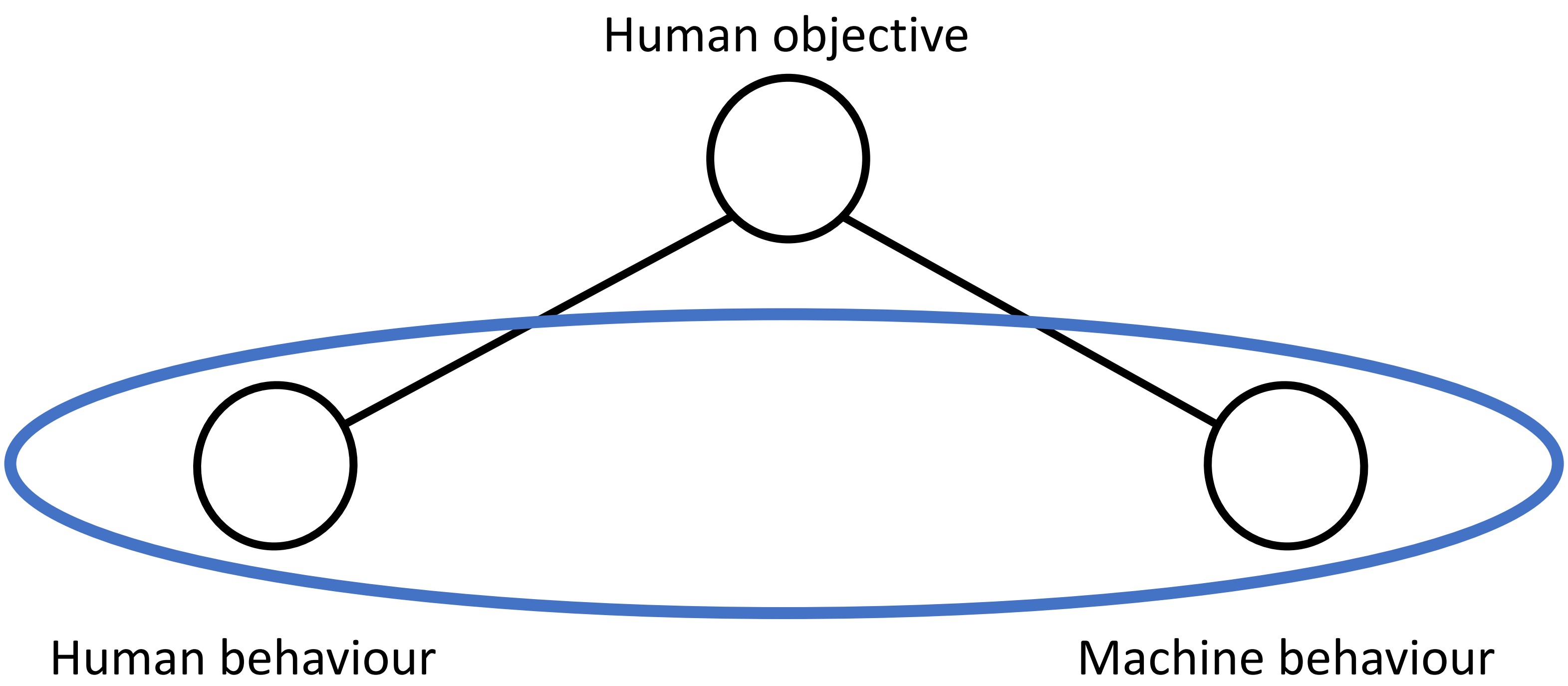
AIMA 1,2,3: objective given to machine



AIMA 1,2,3: objective given to machine

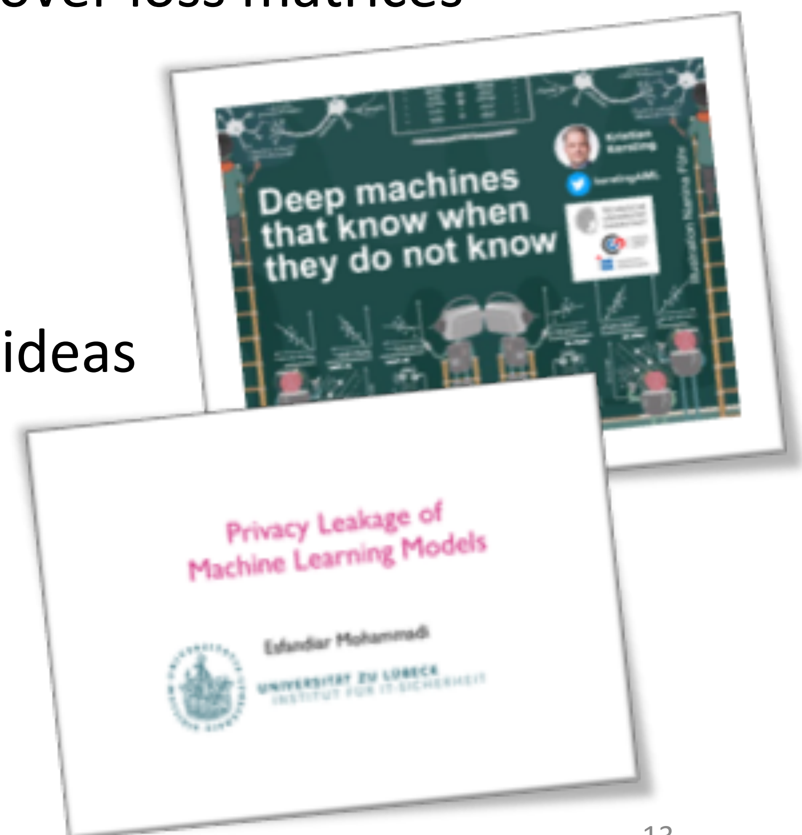


AIMA 4: objective is a latent variable



Example: image classification

- Old: minimize loss with (typically) a *uniform* loss matrix
  - Accidentally classify human as gorilla
  - Spend millions fixing public relations disaster
- New: structured prior distribution over loss matrices
  - Some examples safe to classify
  - Say “don’t know” for others
  - Use active learning to gain additional feedback from humans
- Other researchers work on similar ideas
  - E.g., Kristian Kersting
- Sometimes in conflict with demands of privacy
  - E.g., Esfandiar Mohammadi



https://www.ml.informatik.tu-darmstadt.de/papers/waterloo2019talk.pdf
https://www.ifis.uni-luebeck.de/~moeller/KI-Kolloquium/2020-01-13-Mohammadi.pdf

Example: fetching the coffee

- What does "fetch some coffee" mean?
- If there is so much uncertainty about preferences, how does the robot do anything useful?
- Answer:
  - The instruction suggests coffee would have higher value than expected a priori, ceteris paribus
  - Uncertainty about the value of other aspects of environment state doesn't matter *as long as the robot leaves them unchanged*

Basic assistance game

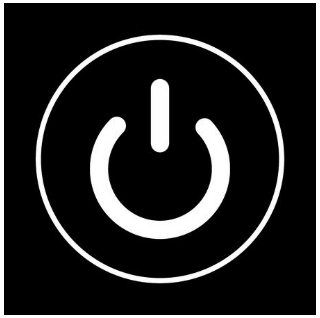
Preferences θ
Acts roughly according to θ

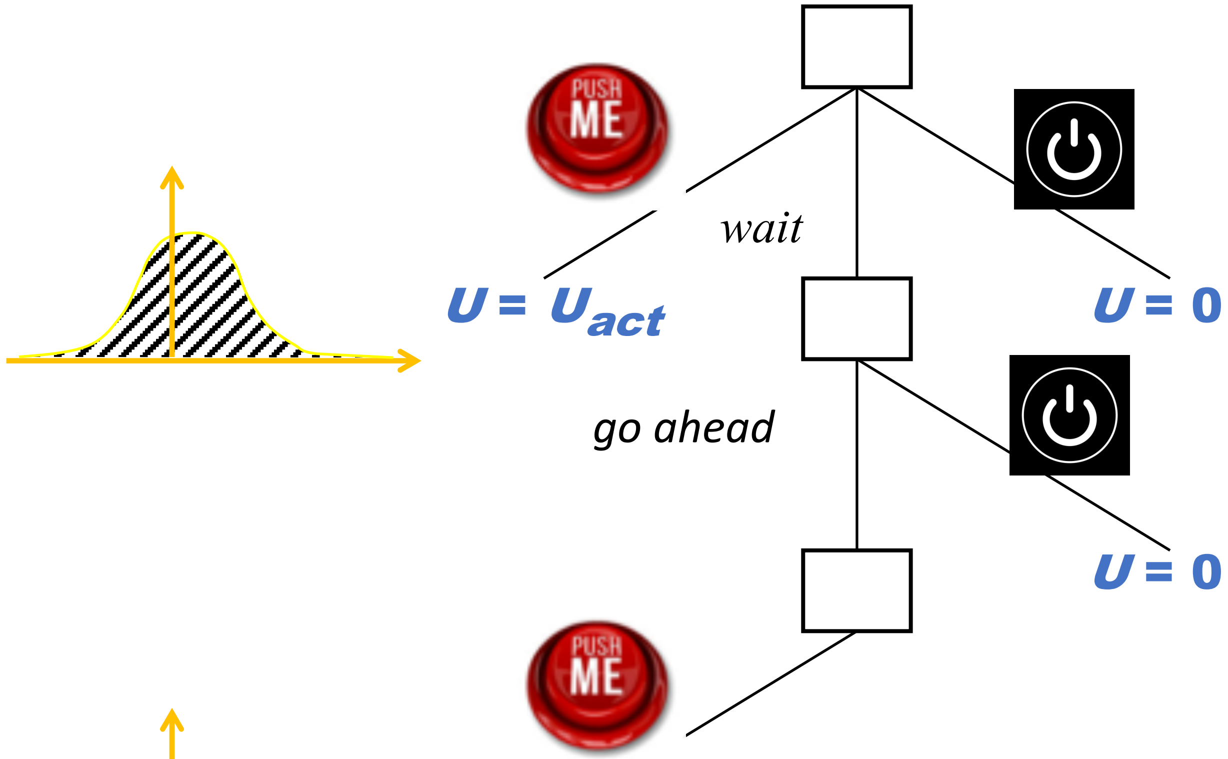
Maximise unknown human θ
Prior P(θ)

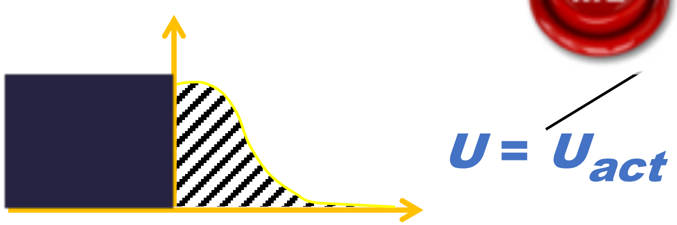
Equilibria:
Human teaches robot
Robot learns, asks questions, permission; defers to human; allows off-switch
Related to inverse RL, but two-way

The off-switch problem

- A robot, given an objective, has an incentive to disable its own off-switch
  - “You can’t fetch the coffee if you’re dead”
- A robot with uncertainty about objective won’t behave this way







Theorem: ***robot has a positive incentive to allow itself to be switched off***

Theorem: ***robot is provably beneficial***

Summary

- Provably beneficial AI is possible ***<u>and desirable</u>***

<u>It isn't "AI safety" or "AI Ethics," it's AI</u>

- Continuing theoretical work (AI, CS, economics)
- Initiating practical work (assistants, robots, cars)
- Inverting human cognition (AI, cogsci, psychology)
- Long-term goals (AI, philosophy, polisci, sociology)

Outline

Provably beneficial AI (Russell)

- Motivation
- Modelling formalism

Human-aware decision making (Rao et al.)

- Mental models
- Interpretable Behaviour
- Explanations

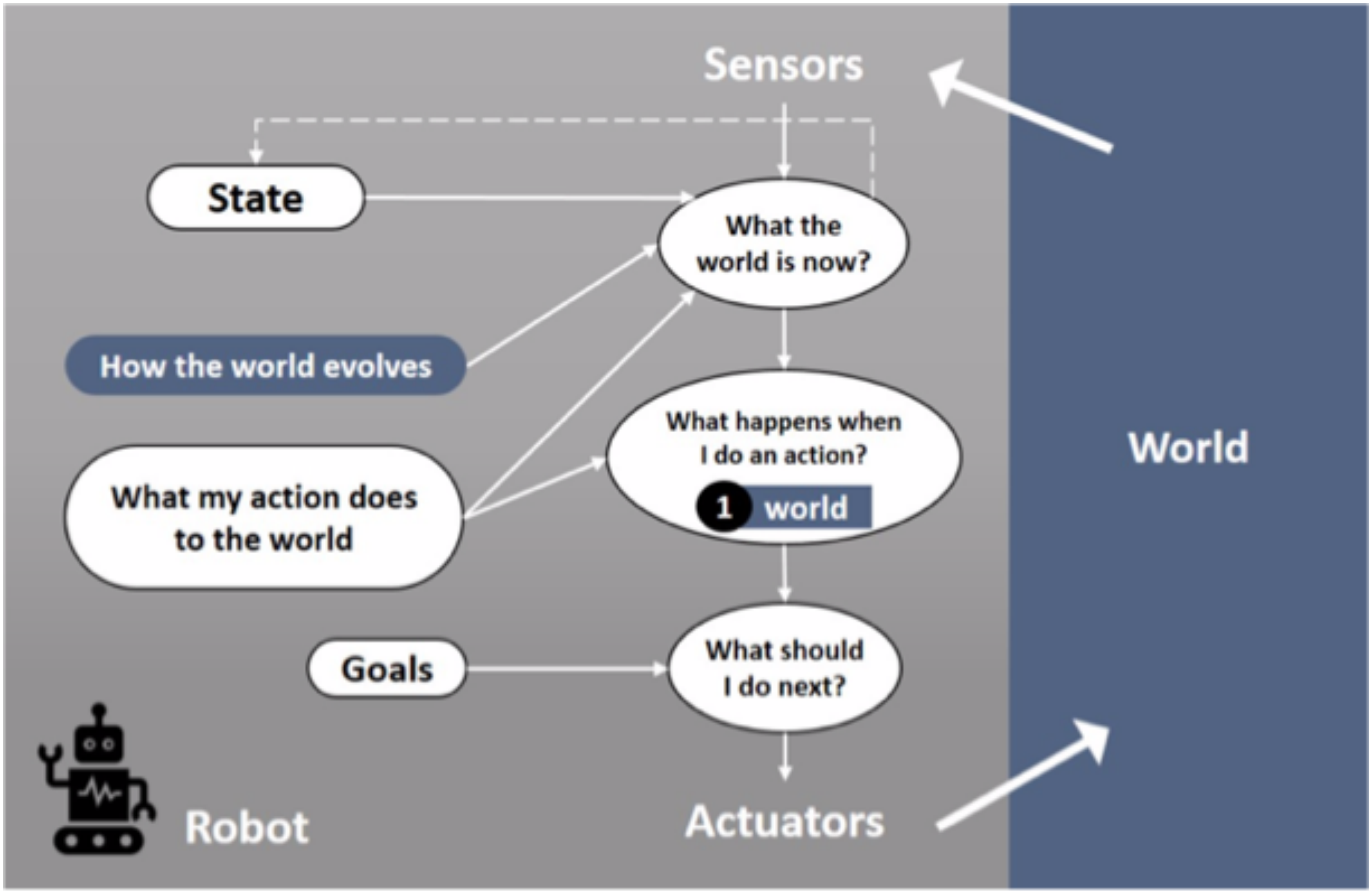
Motivation

- Collaborations between people and AI systems
  - I.e., systems with humans in the loop
  - Augment perception, cognition, problem-solving abilities of people
  - Examples
    - Help physicians make more timely and accurate diagnoses
    - Assistance provided to drivers of cars to help them avoid dangerous situations and crashes
- Objective: Systems that can interact intuitively with users and enable seamless machine-human collaborations
  - Explainable behaviour
    - Explainable AI = XAI

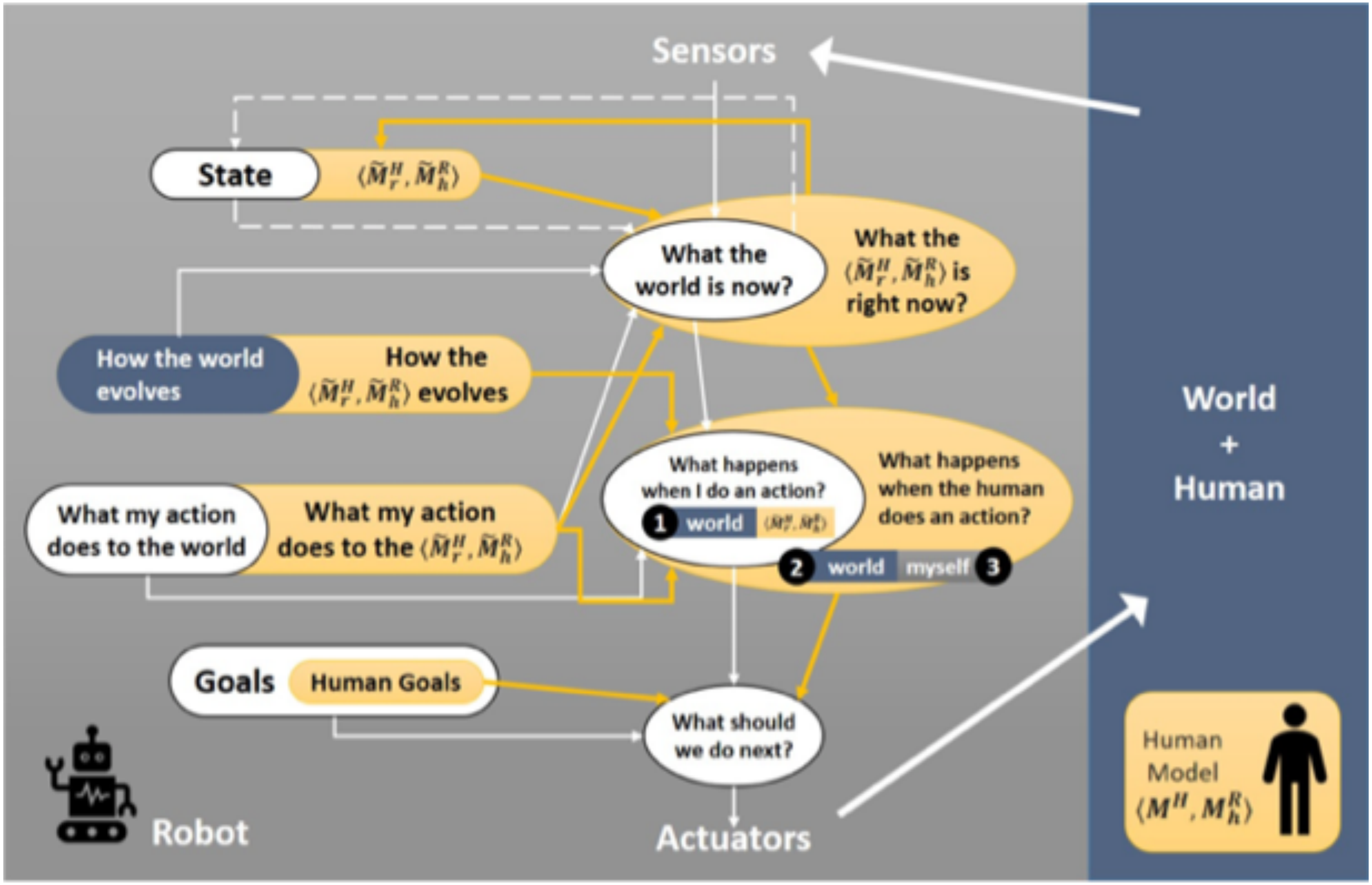
Proposed Solution

- Goal: Synthesise explainable behaviour
- Take into account the mental model of the human in the loop
    - Mental model:
        - Goals + capabilities of the humans in the loop
        - Human's model of AI agent's goals + capabilities

Classical Intelligent Agent

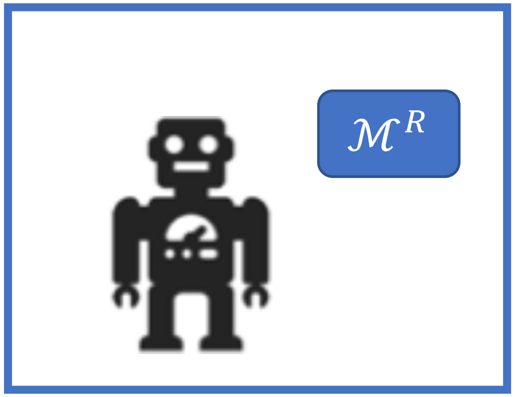


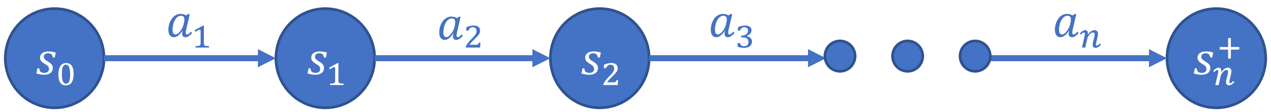
Human-aware Intelligent Agent



Classical Planning

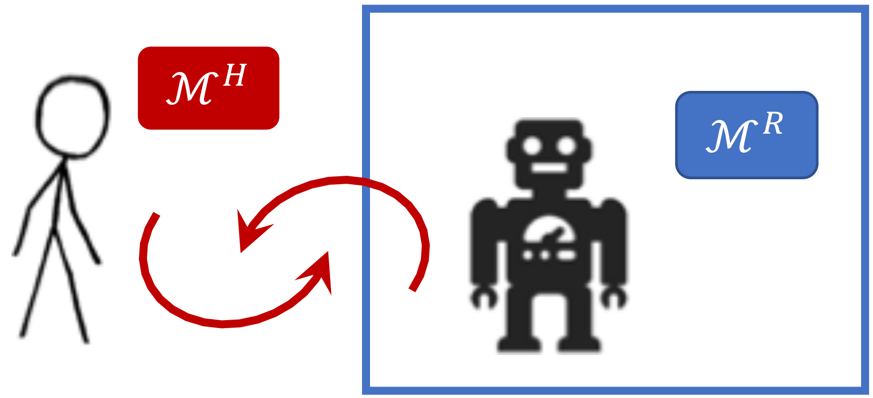
- Given (Σ, s_0, S_g), i.e., the agent's model $\mathcal{M}^R$
- Find a plan $\pi = \langle a_1, a_2, \dots, a_n \rangle$ that transforms s_0 to a state $s_n \in S_g$

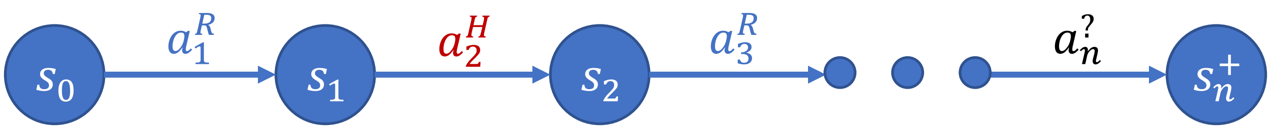




Collaborative Planning

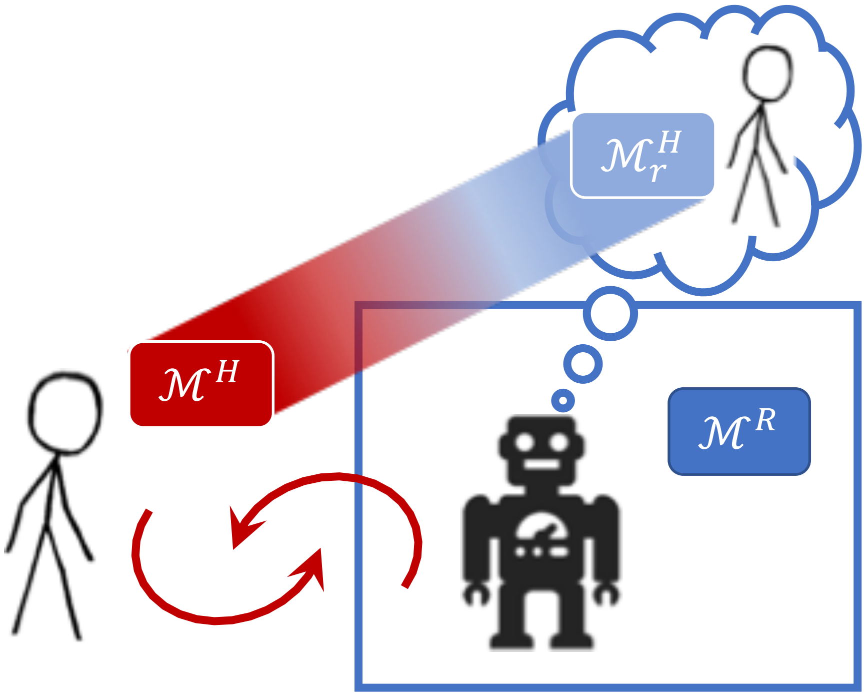
- Given (Σ, s_0, S_g), i.e., the agent's model $\mathcal{M}^R$
- Find a joint plan $\pi = \langle a_1^R, a_2^H, \dots, a_n^? \rangle$ that transforms s_0 to a state $s_n^+ \in S_g$

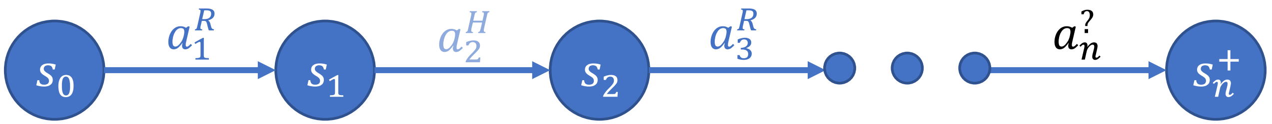




Human-aware Planning

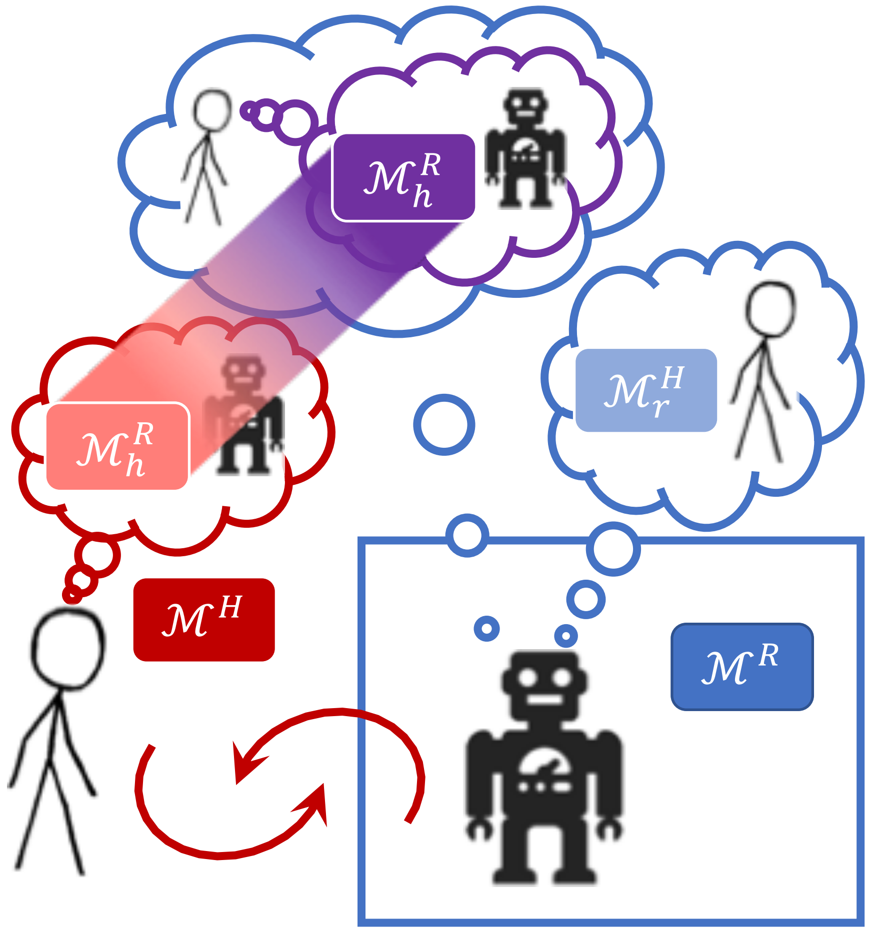
- Next to $\mathcal{M}^R$
- Agent's model $\mathcal{M}_r^H$ of the human's model $\mathcal{M}^H$
  - Allows the agent to **anticipate** human behaviour to
    - assist
    - avoid
    - team

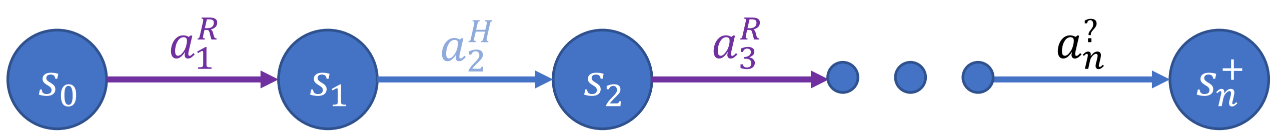




Human-aware Planning

- Next to $\mathcal{M}^R$ and $\mathcal{M}_r^H$
- Agent's model $\widetilde{\mathcal{M}}_h^R$ that the agent expects the human to have of $\mathcal{M}^R$
  - Allows the agent to **anticipate** human expectations to
    - conform to those expectations
    - explain its own behaviour in terms of those expectations





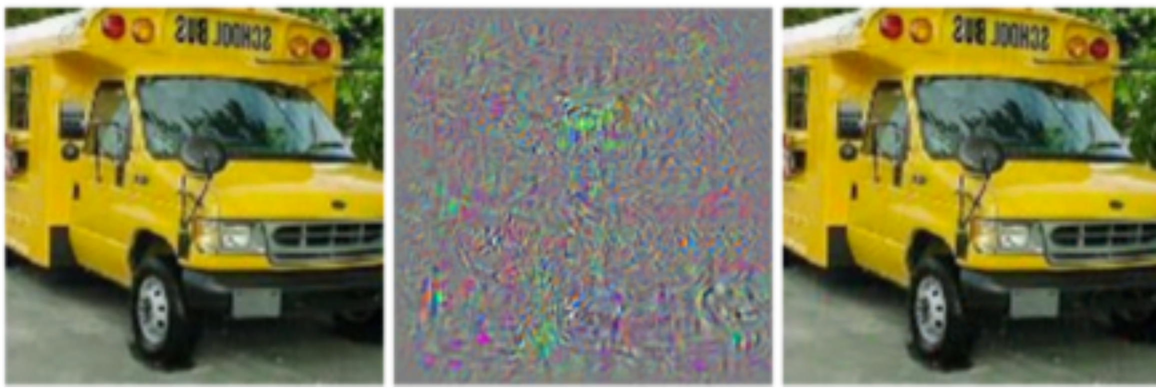
Generating Mental Models

- Known beforehand (handcrafted/researched)
  - Urban Search and Rescue
  - Teaching
- Learning simple models for generating explanations/explicability
- Learning full models (transition functions, rewards)
  - Through interaction with users

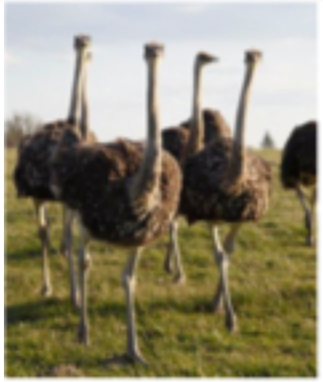
XAI & Explanations

- Standard XAI: view of explanations too simple
  - Debugging tool for “inscrutable” representations
    - “Pointing” explanations (primitive)

Please point to the “ostrich” part



Prediction: School bus | Difference between left and right magnified by 10 | Prediction: Ostrich



    - Explaining decisions will involve pointing over space-time tubes
- Explanations critical for collaboration
  - But not as a monologue from the agent → interaction

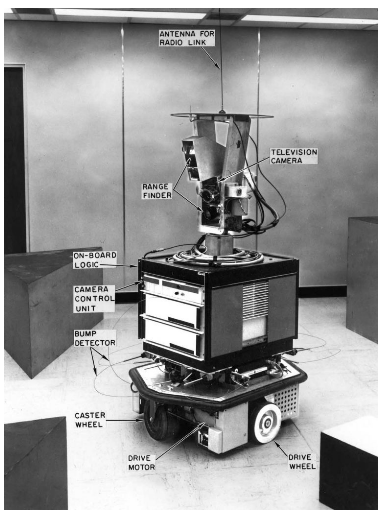
Ethical Quandaries of Interaction

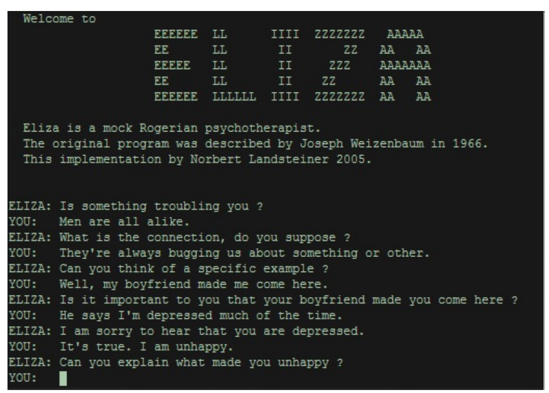
- Evolutionary, mental modelling allowed us to both cooperate or compete/sabotage each other
  - Lying is only possible because we can model others' mental states
- Human-aware AI systems with mental modelling capabilities bring additional ethical quandaries
  - E.g., automated negotiating agents that misrepresent their intentions to gain material advantage
  - Your personal assistant that tells you white lies to get you to eat healthy (or not…)

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

Ethical Quandaries of Interaction

- Humans' example closure tendencies are more pronounced for emotional/social intelligence aspects
  - No on who saw Shakey the first time thought it could shoot hoops, yet the first people interacting with Eliza assumed it was a real doctor
  - Concerns about human-aware AI "toys" such as Cozmo (e.g., Sherry Turkle)





https://thenewstack.io/remembering-shakey-first-intelligent-robot/
https://en.wikipedia.org/wiki/ELIZA https://anki.com/en-us/cozmo.html

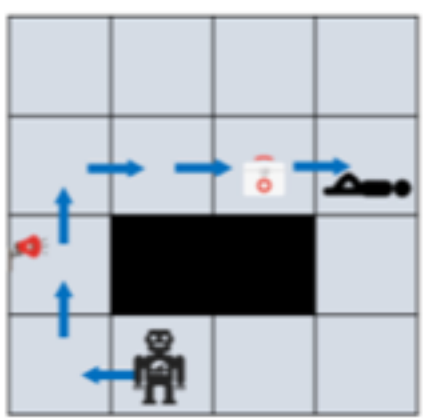
Differences in Mental Models

- Expectations on capabilities
  - Human may have misconceptions about robot's actions
  - Certain actions in human's mental model may not be feasible for robot
- Expected state of the world
  - Human may assume certain facts are true (when they are not true)
- Expected goals
  - Human may have misconceptions about robot's objectives/intentions
- Sensor model differences
  - Human may have partial observability of robot's activities
  - Human may have incorrect beliefs about robot's observational capabilities
- Different representations
  - Robot's innate representation scheme might be too complex for human
  - Human may be thinking in terms of a different vocabulary

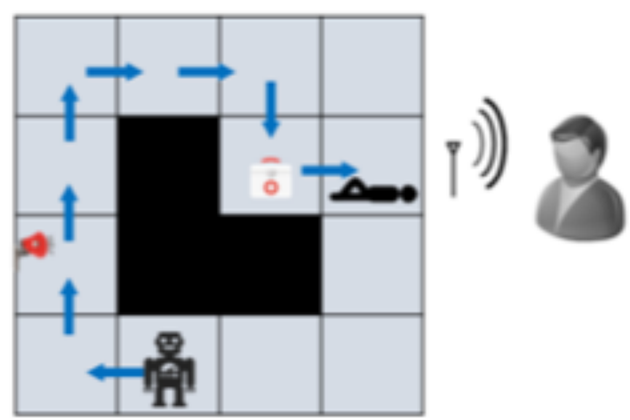
Urban Search and Rescue (USAR)

- Robot deployed to a disaster area
- Tasks robot can perform
  - Survey particular rooms
  - Identify survivors
  - Perform triage
- Two agents in domain
  - Internal agent – Robot
  - External agent – Human
- Their models may diverge – leading to different expectation on behaviours

Robot's model



Human's mental model of the Robot Model



Model Differences

- Robot and human may have different models of same task
  - Divergence in models can lead to expectation mismatch
  - Consequence: Plans that are optimal to robot may not be so in model of human
    - Inexplicable plans
- Robot has two options
  - Explicable planning – sacrifice optimality in own model to be explicable to human
    → *interpretable behaviour*
  - Plan Explanations – resolve perceived suboptimality by revealing relevant model differences
    → *model reconciliation*

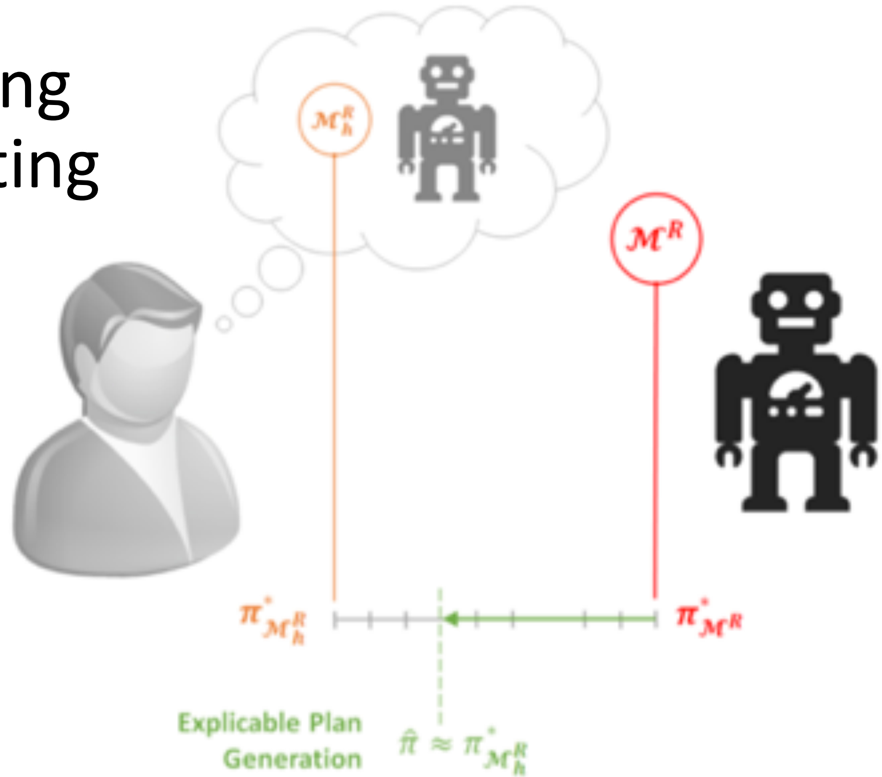
Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Interpretable Behaviour

- Explicable behaviour
  - Acting in a way that make sense to the user
- Legible behaviour
  - Acting in a way that convey necessary information to the user
- Predictable behaviour
  - Acting in a way that allow users to accurately anticipate future behaviour

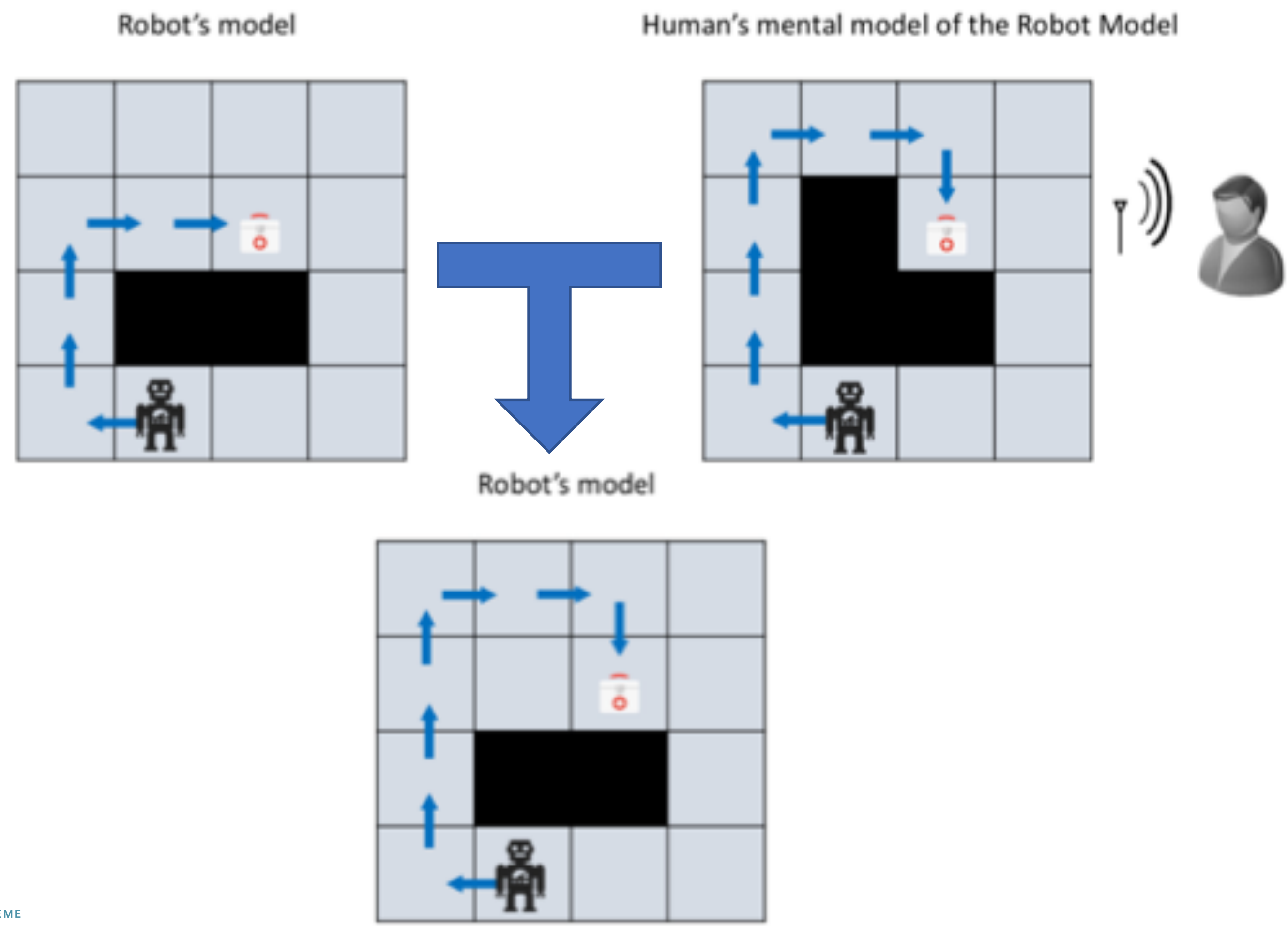
Explicable Behaviour

- Robot's behaviour may diverge from human's expectations of it
- Human may get surprised by robot's inexplicable behaviour
- One way to avoid surprising a human involves generating explicable behaviour by conforming to human's expectations
  - Account for human's mental model



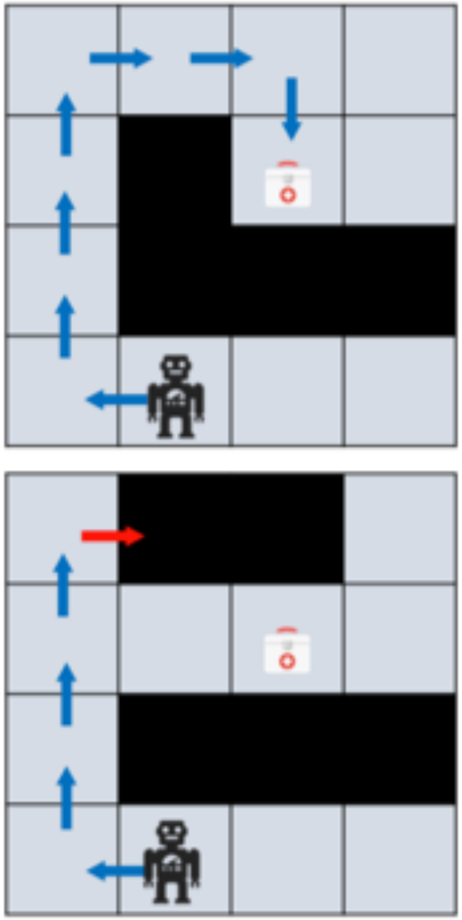
Explicable Behaviour

- Example: Robot may have to sacrifice its optimality to improve explicability



Model-based Explicable Behaviour

- Human's mental model is available to the robot
- Robot cannot plan directly with human mental model
- Find a valid plan that is 'closest' to the expected plan
- Involves minimizing distance w.r.t. expected plans
  - Cost difference in human model
  - Action set difference



Kulkarni A, Zha Y, Chakraborti T, Vadlamudi SG, Zhang Y, Kambhampati S. Explicable planning as minimizing distance from expected behavior. In Proceedings of AAMAS 2019 as extended abstract

Model-free Explicable Planing

- Human's mental model may not be known upfront

$$\underset{\pi_{\mathcal{M}^R}}{\operatorname{argmin}} \boxed{cost(\pi_{\mathcal{M}^R})} + \alpha \cdot \boxed{dist\left(\pi_{\mathcal{M}^R}, \pi_{\mathcal{M}_h^R}\right)}$$

Cost of robot plan

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

- We do not necessarily need to learn the full model

Zhang Y, Sreedharan S, Kulkarni A, Chakraborti T, Zhuo HH, Kambhampati S. Plan explicability and predictability for robot task planning. In Proceedings of ICRA 2017

Model-free Explicable Planing

- Understand = Associate abstract tasks with actions
- Consider as a labelling process

$$\underset{\pi_{\mathcal{M}^R}}{\operatorname{argmin}}\, cost(\pi_{\mathcal{M}^R}) + \alpha \cdot \boxed{dist\left(\pi_{\mathcal{M}^R}, \pi_{\mathcal{M}_h^R}\right)}$$

$$\underbrace{\boxed{F} \circ \boxed{\mathcal{L}_h(\pi_{\mathcal{M}_R})}}$$

Domain-independent function taking task labels as inputs, returning approx. distance value

Labelling scheme of human for agent plans (to be learned)

Input

Function composition

Input

F = ($task_1$, $task_2$, $task_3$)

E.g., the ratio between number of actions with non-empty labels and the number of all actions

Output

Plan = {a_1, a_2, a_3, …, a_n}

L_h = $task_1$ $\perp$ $task_2$ $task_1$

No label – inexplicable

Output

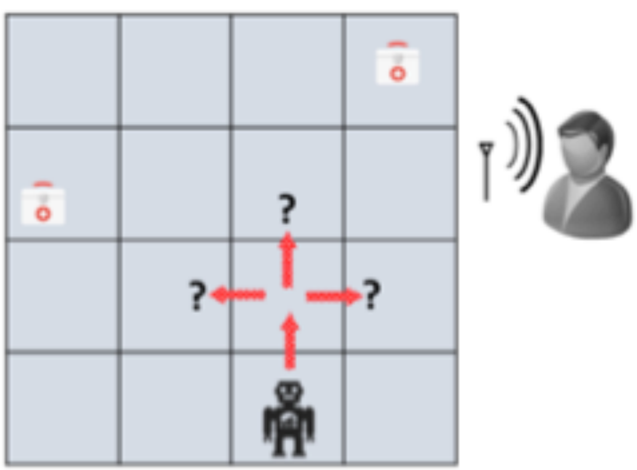
Zhang Y, Sreedharan S, Kulkarni A, Chakraborti T, Zhuo HH, Kambhampati S. Plan explicability and predictability for robot task planning. In Proceedings of ICRA 2017

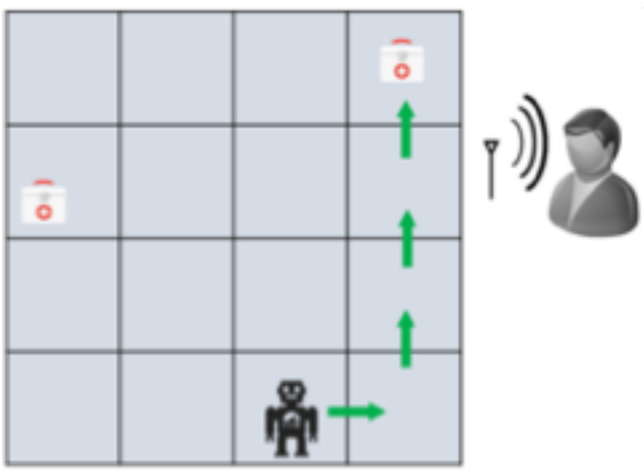
Why **Legible** Behaviour?

- In human-robot teams, essential for the robot to communicate its intentions and objectives to the human
  - Explicitly communicate its intentions to the human
  - Generating a behaviour which implicitly reveals robot's intentions to the human
    - Might be easier for the human teammate

Legible Behaviour

- In general, involves a setting where
  - Human has access to candidate goals but does not know true goal
- Robot's objective: Convey true goal implicitly through its behaviour
- Human updates its belief on set of candidate goals when it receives observations
- By synthesizing legible behaviour, robot reduces human's uncertainty over candidate goals





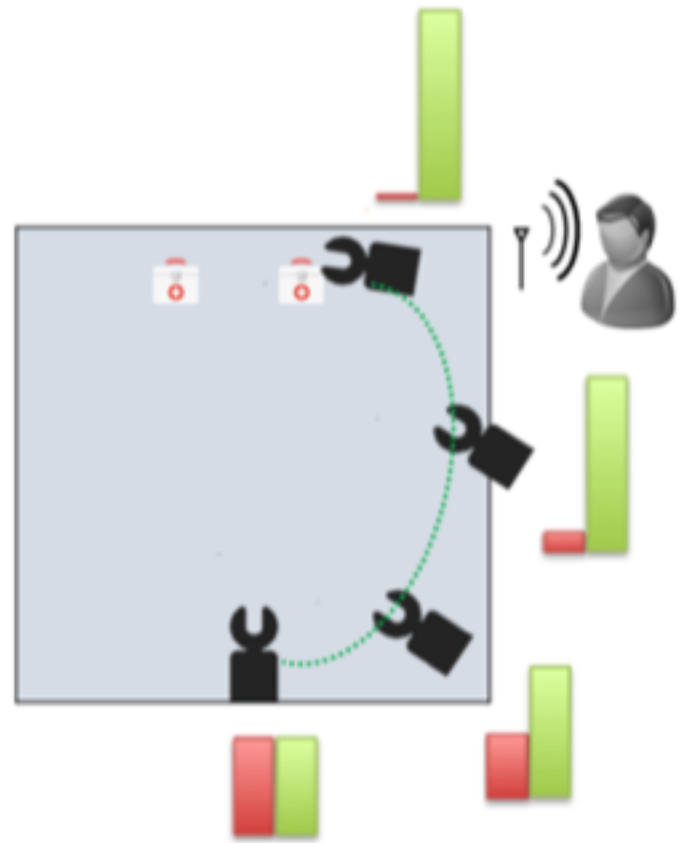
Online Legible Behaviour

- Enables human to quickly and confidently infer robot's true goal
- Human's belief update is captured using a probabilistic goal recognition system
- Actions that maximize the posterior probability of the true goal G are favoured

$$\operatorname*{argmax}_{G \in \mathcal{G}} P(G|Observations)$$

Legible Robot Motion

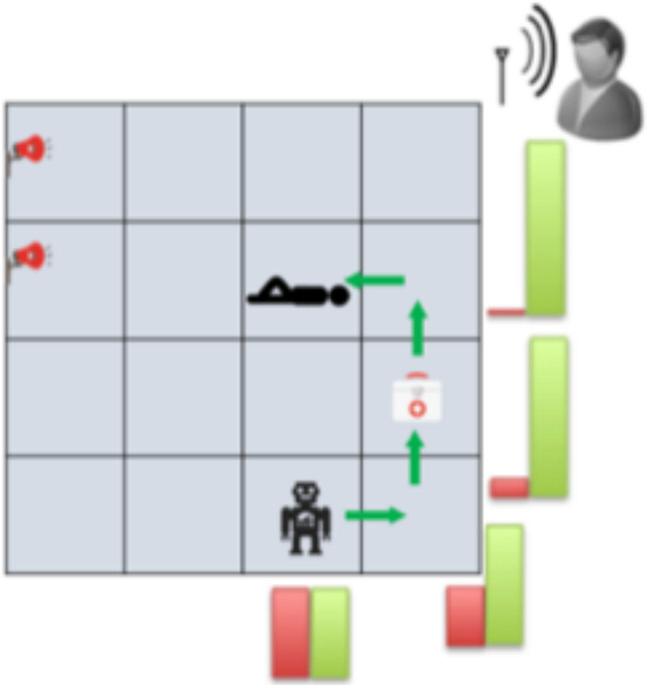
- Example: Which medkit will the robot pick up?
- While performing goal recognition, human considers shortest distances
- Approach involves finding a trajectory endpoint between start point and true goal such that posterior probability of true goal is maximized
  - Sooner the goal is recognized in the trajectory, the better is the trajectory's legibility



Dragan AD, Lee KC, Srinivasa SS. Legibility and predictability of robot motion. In Proceedings of HRI 2013.

Transparent Planning

- Example: Is the robot surveying the rooms or performing triage?
- Whenever an action is performed, goal recognition system is used to update human's belief
- Objective: Reach a target belief where true goal is more probable than other goals
- Take the first applicable action associated with a belief of highest utility (closest to target belief)



MacNally AM, Lipovetzky N, Ramirez M, Pearce AR. Action selection for transparent planning. In Proceedings of AAMAS 2018.

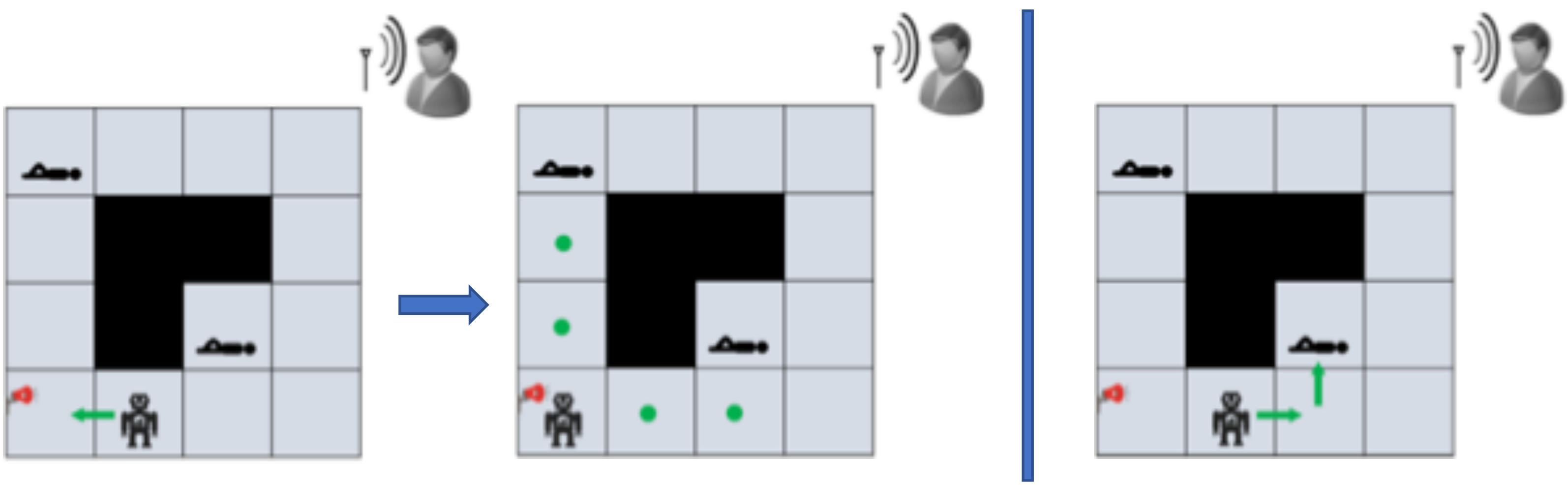
Offline Goal Legibility

- Generalizes problem of goal legibility in terms of
  - Partial observability of the human
  - Amount of goal legibility achieved
- Partial observability:
  - Multiple action and state pairs may yield the same observation
  - Human's belief update consists of all possible states that emit given observation and are valid considering previous belief
  - $b_{i+1} = update(b_i, o_{i+1})$

Kulkarni A, Srivastava S, Kambhampati S. A unified framework for planning in adversarial and cooperative environments. In Proceedings of the AAAI 2019.

Offline Goal Legibility

- Example: Robot has to survey and treat a victim
  - Has to convey which victim it is treating
- Key idea: Limit number of candidate goals (at most j goals) possible in observer's final belief
- Explores legible behaviour that satisfies predetermined amount of goal legibility, i.e., the plan is j-legible



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR INFORMATIONSSYSTEME

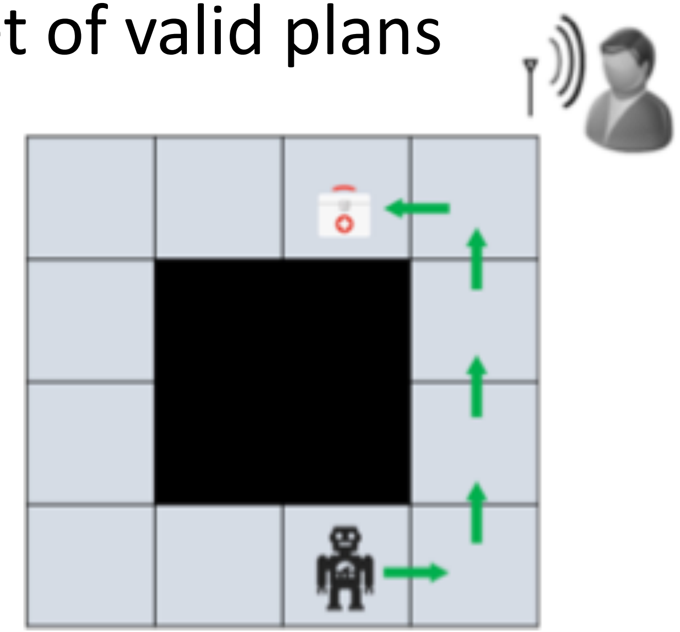
Kulkarni A, Srivastava S, Kambhampati S. A unified framework for planning in adversarial and cooperative environments. In Proceedings of the AAAI 2019.

Why **Predictable** Behaviour?

- In human-robot teams, if robot's behaviour cannot be anticipated by human, it can hamper team performance
- Predictable robot behaviours are easy for the human to understand and help in engendering trust in the robot
- *Predictability and legibility are fundamentally different and often contradictory properties of motion*

Predictable Behaviour

- In general, involves a setting where
  - Human knows start state and goal but does not know which plan will be executed
- Robot's objective is to behave in a way that can be anticipated by the human
- Observer updates its belief on set of valid plans when it receives observations
- By synthesizing predictable behaviour, robot reduces human's uncertainty over possible behaviours

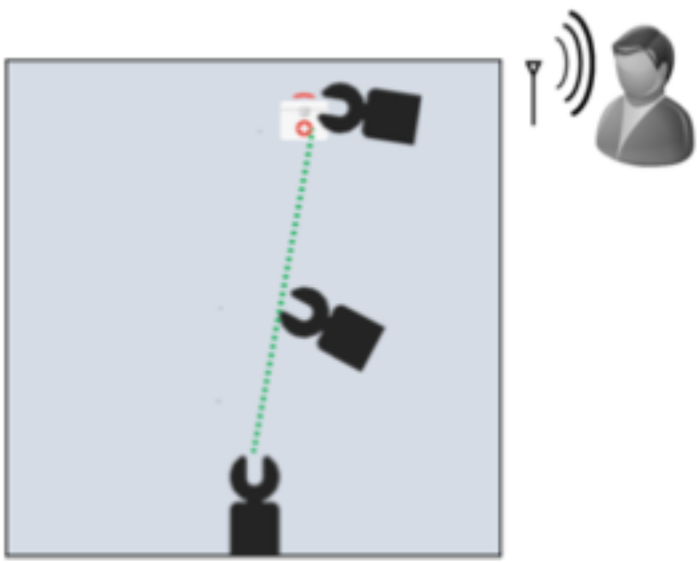


Predictable Robot Motion

- Example: What trajectory will robot take?
- Human assumes that robot is rational and that it prefers short length trajectory
- Most predictable trajectory optimises path towards the goal (C cost fct. modelling human's expectation)

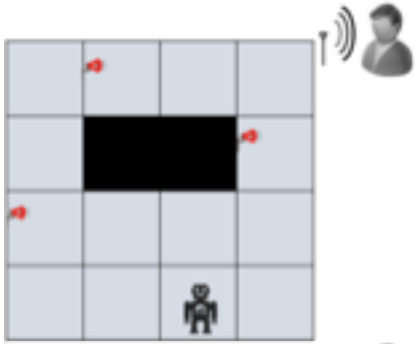
$$\underset{traj}{\operatorname{argmin}}\, C(traj)$$

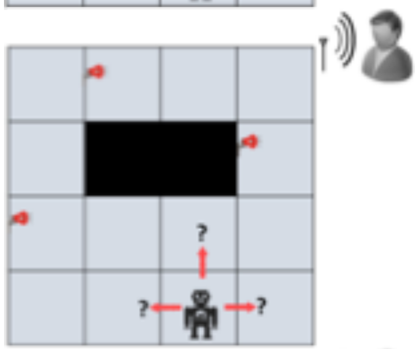
- There are two aspects of generating predictable motion:
  - Learning C
  - Minimizing C

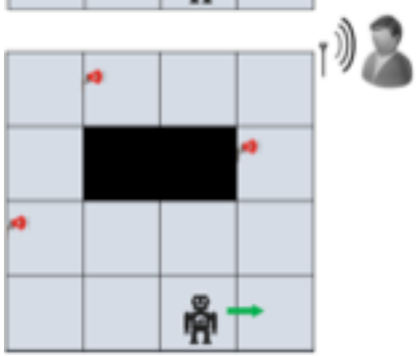


Dragan AD, Lee KC, Srinivasa SS. Legibility and predictability of robot motion. In Proceedings of HRI 2013.

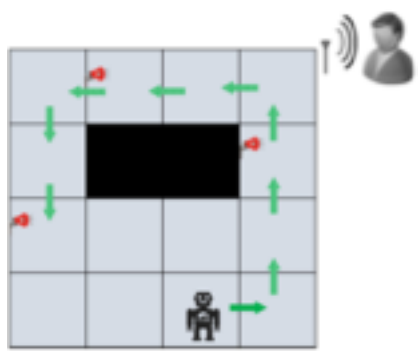
t-Predictability







- Key idea: first t actions should foreshadow rest of actions
- Example: What route would the robot take to survey the rooms?
- t-predictability score P_t = probability of sequence $a_{t+1} \ldots a_T$, given start state, goal and $a_1 \ldots a_t$



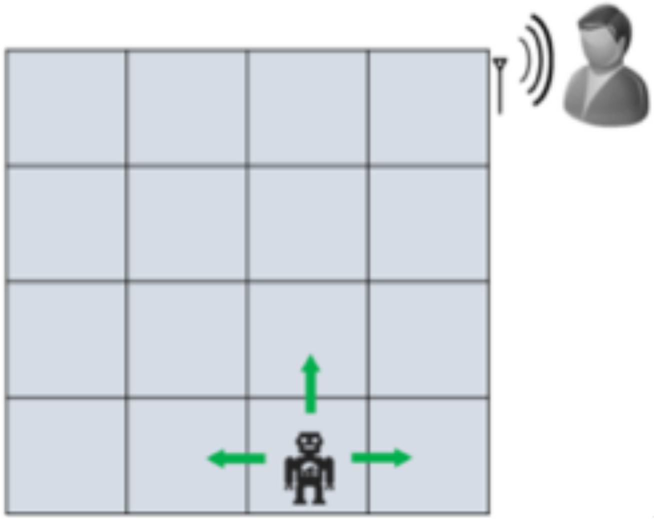
- t-predictable planner finds action sequence $\boldsymbol{a}^*$ such that

$$\boldsymbol{a}^* = \underset{\boldsymbol{a} \in A}{\operatorname{argmax}}\, P_t(\boldsymbol{a})$$

Fisac JF, Liu C, Hamrick JB, Sastry SS, Hedrick JK, Griffiths TL, Dragan AD. Generating plans that predict themselves. In Proceedings of WAFR 2016.

Offline Plan Predictability

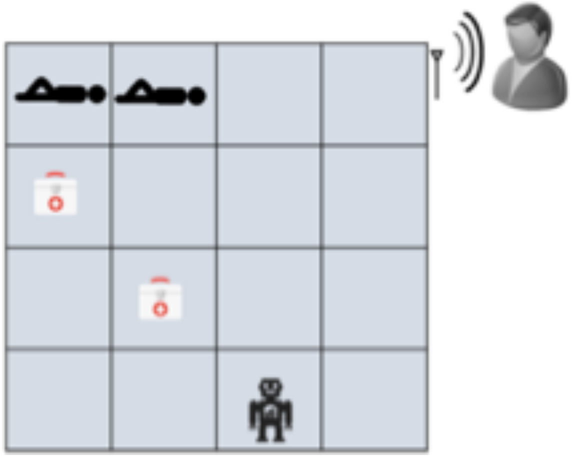
- Assume offline setting
  - Human has partial observability
  - Belief update performed after receiving all observations
- Human guesses robot's actions based on plans that
  - Are consistent with observation sequence
  - Achieve goal
- Generalizes the problem of conveying actions to observer

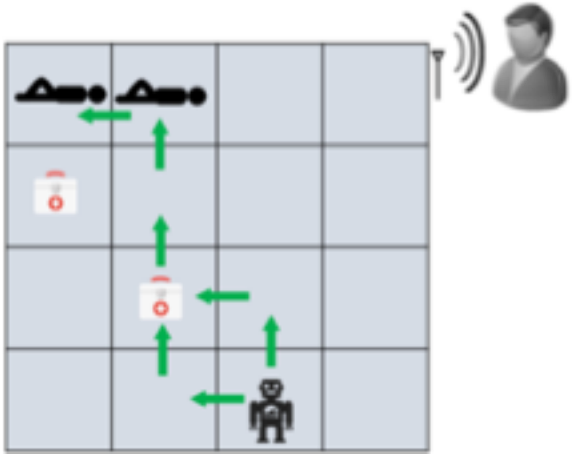


Kulkarni A, Srivastava S, Kambhampati S. A unified framework for planning in adversarial and cooperative environments. In Proceedings of the AAAI 2019.

Offline Plan Predictability

- Example:
  - Robot has to perform triage
  - Which medkit should the robot pick?
- Solution: Generate a plan whose observation sequence is associated with
  - At least m plans to the same goal,
  - And the plans have high similarity.
  - i.e., m plans that are at most d distance from each other – m-similar plans
- Using plan distance metrics
  - Action set distance gives the number of similar actions given two plans



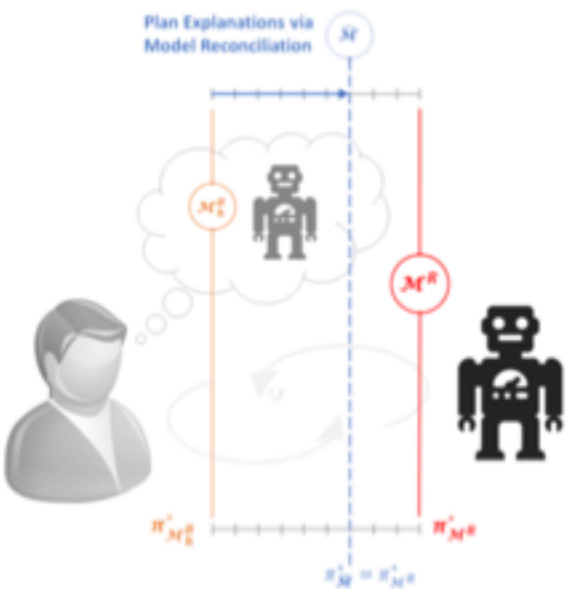


Kulkarni A, Srivastava S, Kambhampati S. A unified framework for planning in adversarial and cooperative environments. In Proceedings of the AAAI 2019.

Plan Explanations

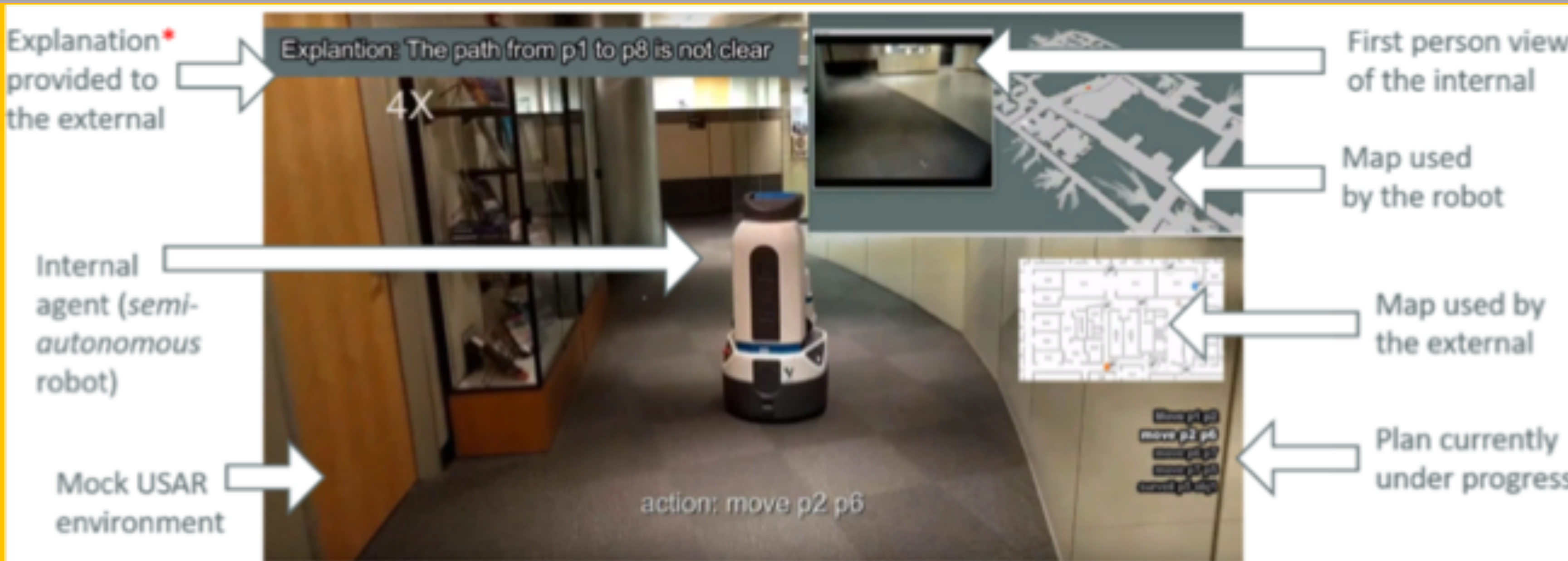
Not in this semester

- Conforming to expectations of human
  - E.g., by explicable planning, considering human's model of the robot as well
  - But: May not be feasible
- Model reconciliation: Bring mental model closer by explanations
  - Planner is optimal in self but not in human's model
  - Given a plan, explanation is a model update
  - After explanation, plan is also optimal in the updated human model



Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Example



- Mock search and reconnaissance scenario with internal robot and external human

Aspects to Explanations

- Completeness: No better explanation exists, no aspect of plan remains unexplicable
  - Requires explanations of a plan to be comparable
- Conciseness: Explanations are easily understandable to the explainee
  - The larger an explanation, the harder for the human to incorporate information into deliberative process
- Monotonicity: Remaining model differences cannot change completeness of explanation, i.e., all aspects of model that yielded plan are reconciled
  - Subsumes completeness
- Computability: Ease of computing explanation from robot's point of view

Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Types of Explanations

- Plan Patch Explanation (PPE)
  - Provide model differences pertaining to only the actions present in the plan that needs to be explained
- Model Patch Explanation (MPE)
  - Provide all model differences to the human
- Minimally Complete Explanation (MCE)
  - Shortest complete explanation
  - Can be rendered invalid given further updates
- Minimally Monotonic Explanation (MME)
  - Shortest explanation preserving monotonicity
  - Not necessarily unique as there may be model differences supporting the same causal links in the plan; exposing one link is enough (to guarantee optimality in the updated model)

Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Aspects of Types of Explanations

- Plan Patch Explanation (PPE)
- Model Patch Explanation (MPE)
- Minimally Complete Explanation (MCE)
- Minimally Monotonic Explanation (MME)

Explanation Type	Completeness	Conciseness	Monotonicity	Computability
PPE	✓	✗	✓	✓
MPE	✗	✓	✗	✓
MCE	✓	✓	✗	?
MME	✓	✓	✓	?

$$|\text{approx.}\, MCE| \leq |\text{exact.}\, MCE| < |MME| \ll |MPE|$$

Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Example – FetchWorld

- Fetch robot whose design requires it to tuck its arms and lower its torso or crouch before moving – not obvious to human navigating

Robot's 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)))
```

Human's Model

```
(: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)))
```

Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Example – FetchWorld

- Initial state and goal: `(: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`

Robot's 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)))
```

Human's Model

```
(: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)))
```


Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Example – FetchWorld

- Robot's optimal plan: `pick-up b1 -> tuck -> move loc1 loc2 -> put-down b1`

Robot's 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

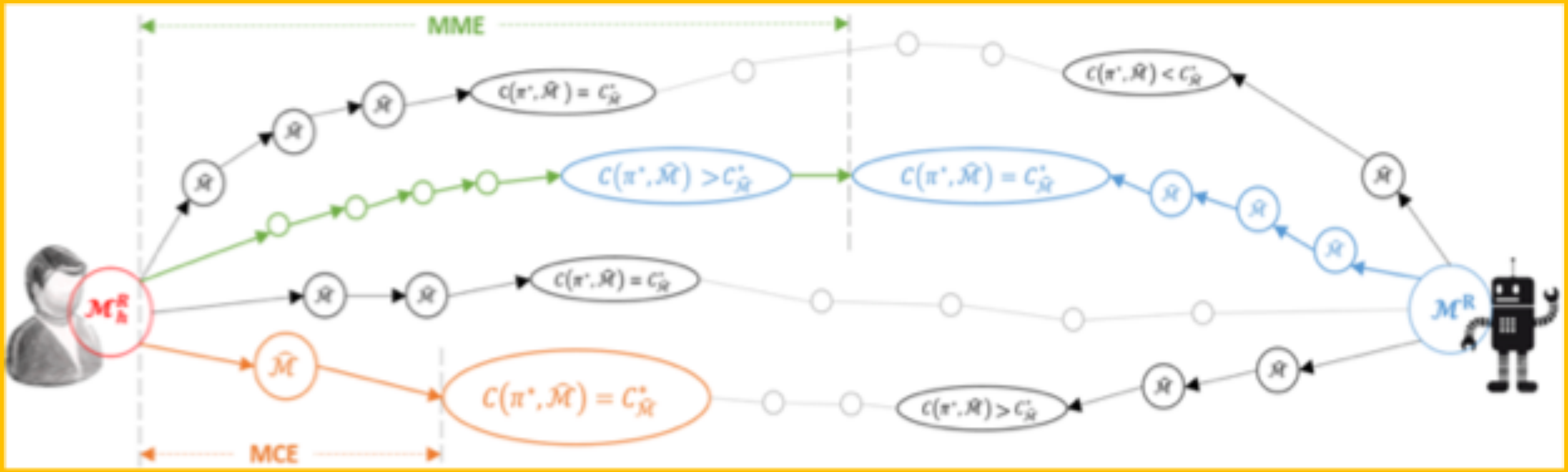
MCE

MME

Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Model Space Search

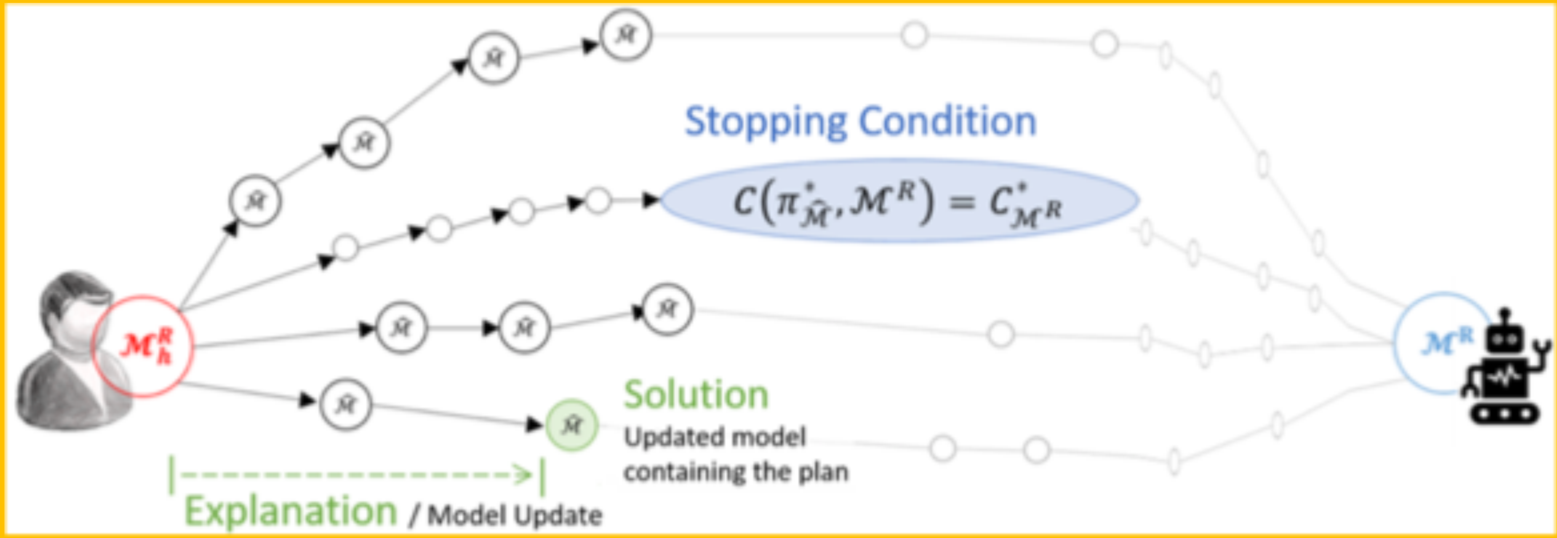
- Search algorithms for finding MCEs and MMEs



Chakraborti T, Sreedharan S, Zhang Y, Kambhampati S. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI 2017.

Model Space Search

- Human-aware planning: Given the model of a planning problem and the mental model of the human, find the right model to plan in
  - Trade-off explicability and explanation

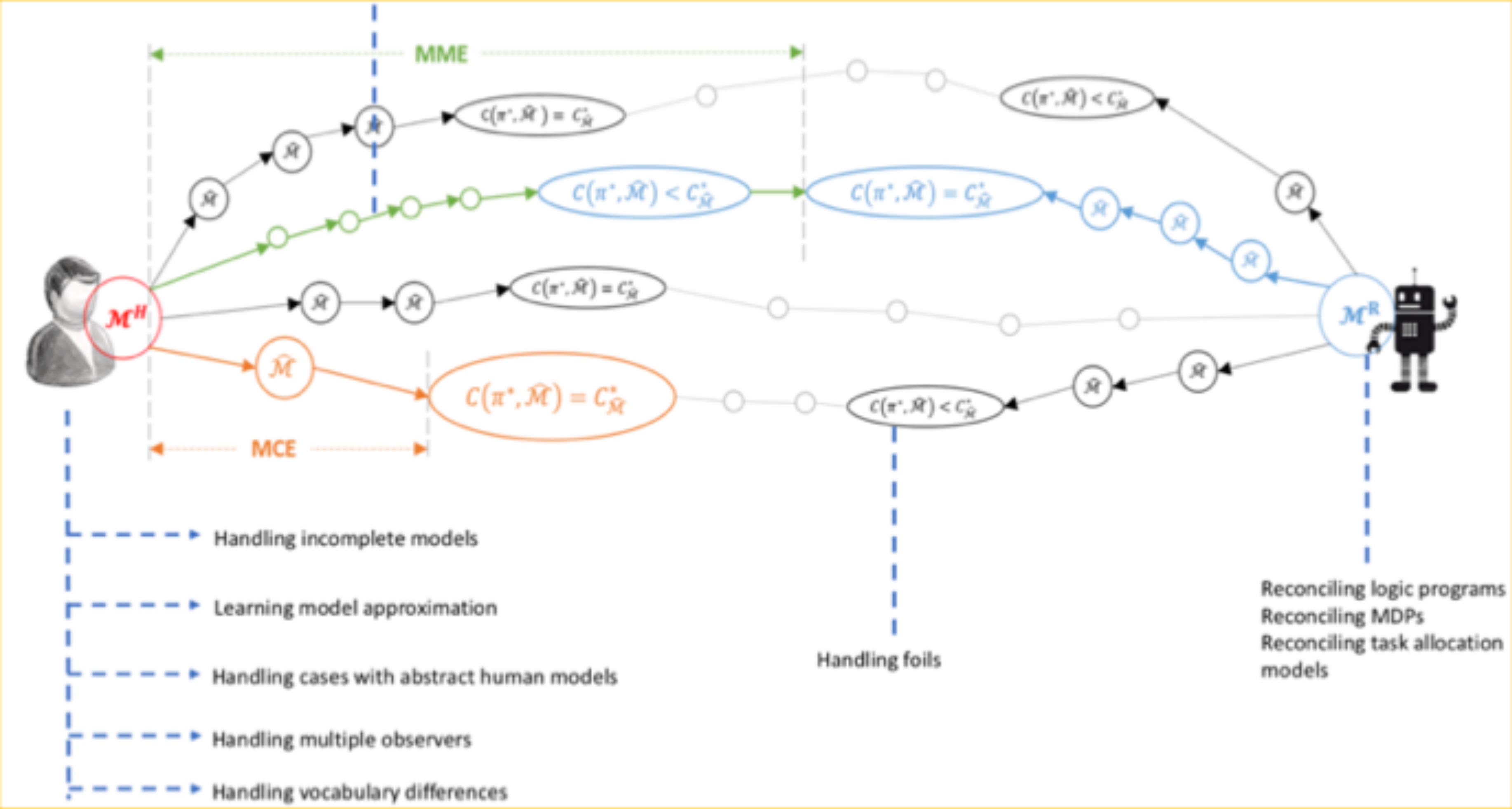


- Minimise
  $cost/length\ of\ explanations + \alpha \cdot departure\ from\ optimality$

Chakraborti T, Sreedharan S, Kambhampati S. Balancing explicability and explanation in human-aware planning. IJCAI 2019.

Extensions (Outlook)

Explanation ordering



Summary

- Mental models
  - Mental model of the human
  - Mental model that the human has of the agent
  - Mental model that the agent assumes the human has of the agent
- Interpretable behaviour
  - Explicability
  - Legibility
  - Predictability
- Explanations (not in this semester)
  - Model reconciliation

Outline

Provably beneficial AI (Russell)

- Motivation
- Modelling formalism

Human-aware decision making (Rao et al.)

- Mental models
- Interpretable Behaviour
- Explanations

The End