On the Computational Complexity of Model Reconciliation

Sarath Sreedharan[1], Pascal Bercher[2] , Subbarao Kambhampati[1]

[1]-Arizona State University

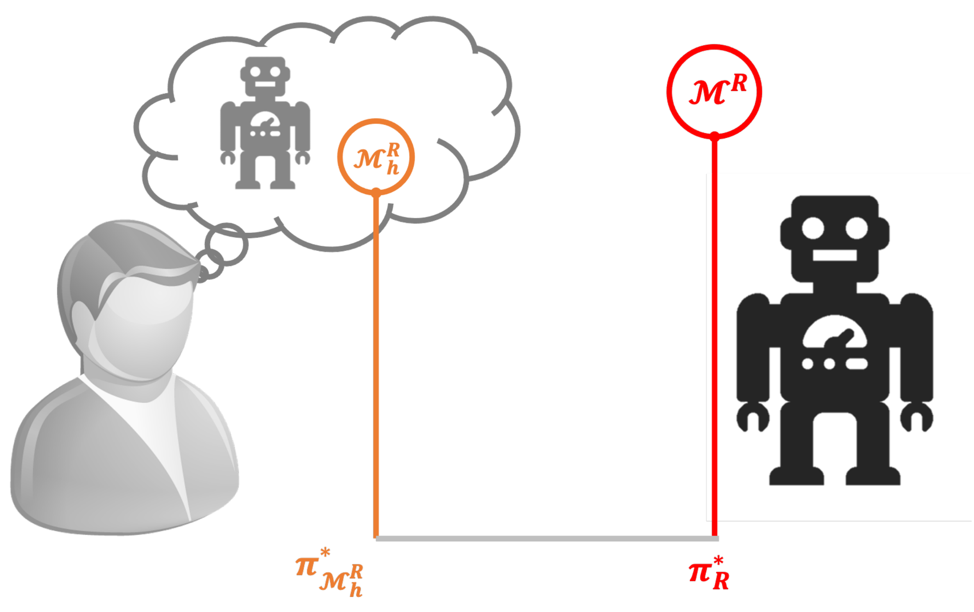
[2]-The Australian National University

Model Reconciliation

Model Reconciliation Problem

$\langle M^R, M_h^R, \pi_R^* \rangle$

M^R - Robot's planning model
M_h^R - The human's belief about the robot model
π_R^* - The plan being proposed by the robot



Human could be confused by the proposed plan, if

$\mathcal{M}_h^R \neq \mathcal{M}^R$

Even if the human is a perfect reasoner π_R^* may be suboptimal or even invalid in $\mathcal{M}_h^R$

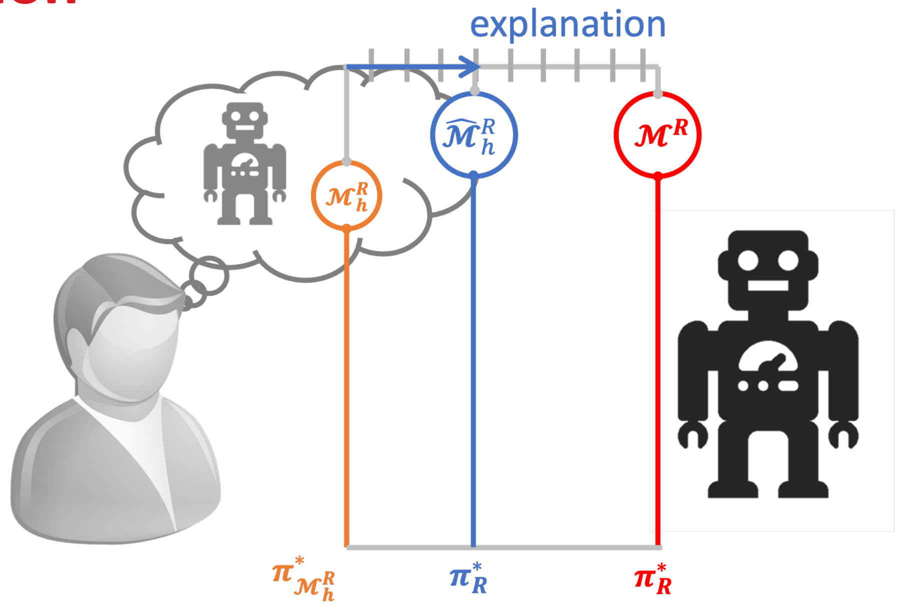
Explanatory Query:
Why did you select π_R^*?

There may be too many differences between the human model and the robot model. Dumping the robot model may overwhelm the user

Model Reconciliation Explanation

$\mathcal{M}_h^R \rightarrow \mathcal{M}^R$

Updated human model after explanation

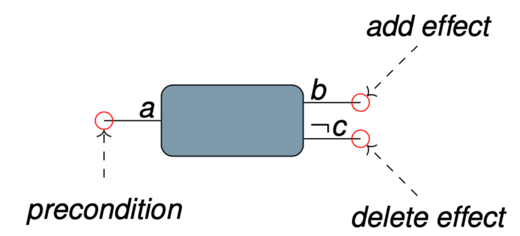


Model reconciliation explanations have generally focused on identifying the minimum number of model updates to be provided to the human so the plan π_R^* will be optimal in the updated model.

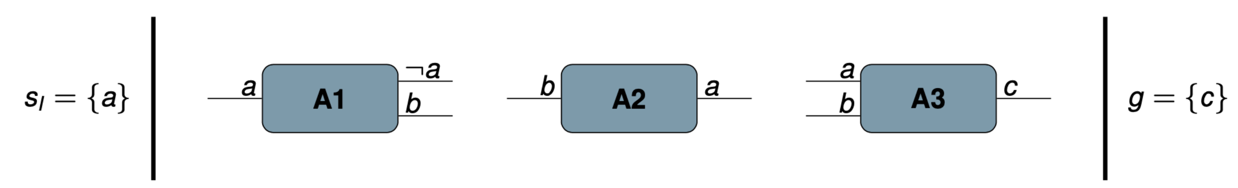
Basic Terminology

In classical planning,

- States are sets of propositional variables *F*
- Actions describe state transitions:

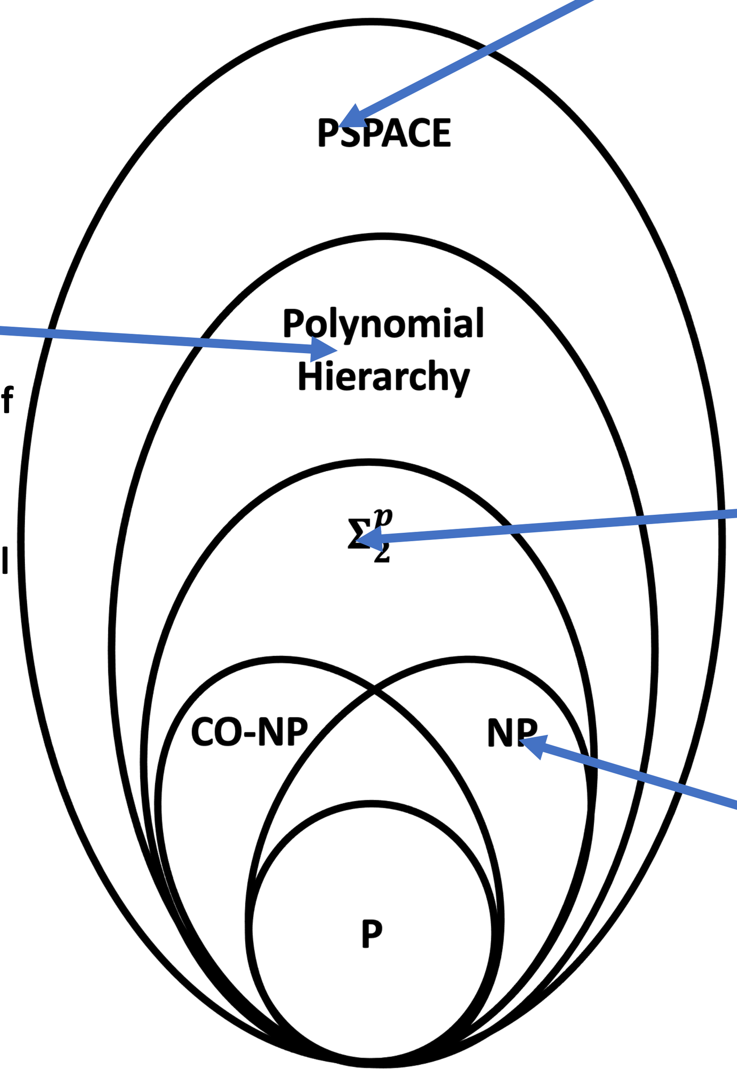


Our goal is to find the right sequence of actions that turns an initial state into a desired (goal) state, e.g.:



Complexity Classes

Complexity for plan existence



Polynomial hierarchy consists of the union of classes of the form Σ_i^p (including Σ_2^p) – Each class Σ_i^p has a canonical problem denoted as $QSAT_i$ containing alternating existential and universal quantifiers

Complexity for ***MRE-k***

Canonical problem: $QSAT_2$

$\exists X\, \forall Y\, \phi(X,Y)$

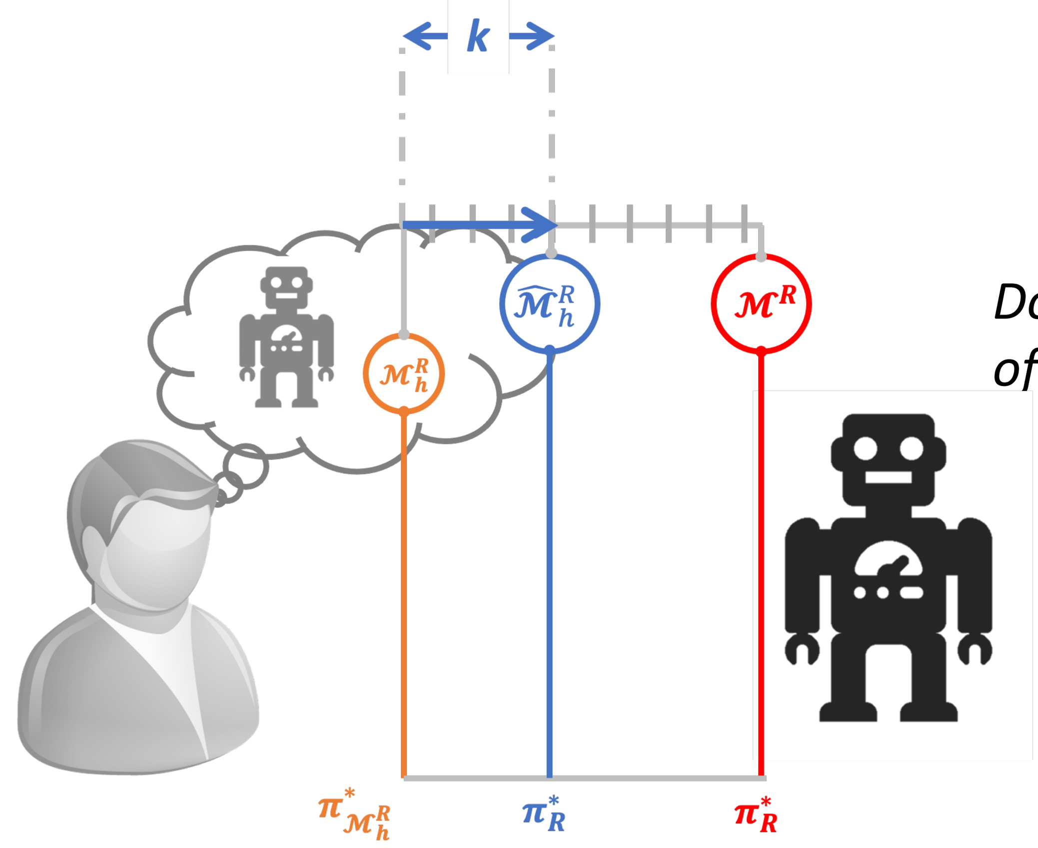
Complexity for bounded plan existence (bound encoded unarily)

In addition to establishing the complexity of model reconciliation explanation generation our result also establishes an alternate method for generating such explanations – namely through QBF compilation

Complexity Results

Proposition 1. The question whether there exists a valid explanation can be decided in constant time. More precisely, the answer is always yes.

Bounded Model Reconciliation Problem (*MRE-k*)



Does there exist a valid explanation of size k?
Where $|\epsilon^+| + |\epsilon^-| = k$

Model updates

MRE-K Complexity

A SAT formula testing whether π_R^* is valid in $M_h^R + \epsilon$

Theorem 1. MRE-k is in Σ_2^p (Membership)

$$\exists\, (X,Z) \forall\, Y\ (\phi_1(X) \wedge \neg(\phi_2(X,Y)) \wedge \phi_3(Z))$$

A set of propositional formulas that correspond to specific k model updates (ϵ) to be applied to M_h^R

A SAT encoding of $M_h^R + \epsilon$ for a planning horizon of $|\pi_R^*| - 1$

$\phi_2(X,Y)$ returns true if there exist a plan of makespan less than $|\pi_R^*|$

Theorem 2. MRE-k is Σ_2^p-hard

$\exists X$ Encoded as possible model updates over initial states

$\forall Y\ \phi(X,Y) \rightarrow \neg(\exists Y\ \neg\phi(X,Y))$

Encoded into an optimality check for π_R^*
- The goal is $\neg\phi(X,Y)$ and possible plans of length $< |\pi_R^*|$ corresponds to various assignments over Y

Theorem 3. MRE-k is Σ_2^p-Complete

Acknowledgement: This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1- 2442, N00014-18-1- 2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF19-2-0006 and a JP Morgan AI Faculty Research grant.