On the Computational Complexity of Model Reconciliation

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

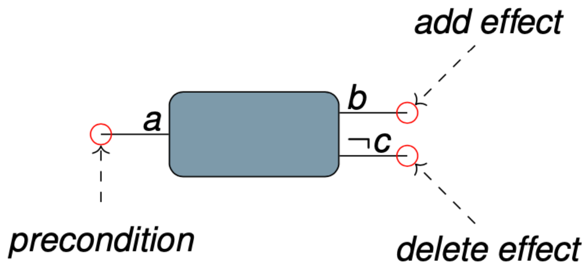
[1]-Arizona State University

[2]-The Australian National University

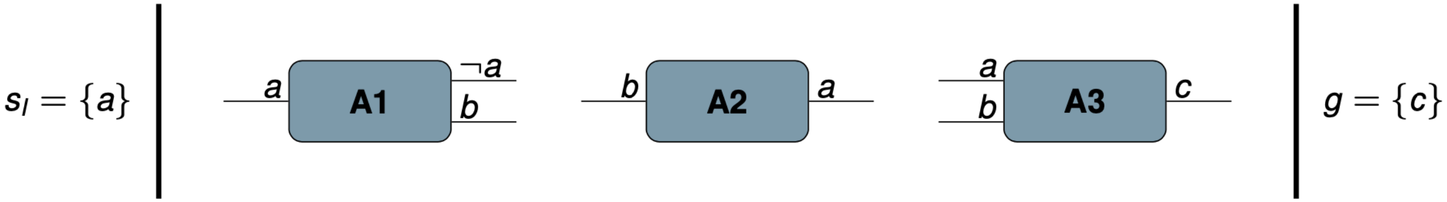
Basic Terminologies

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.:

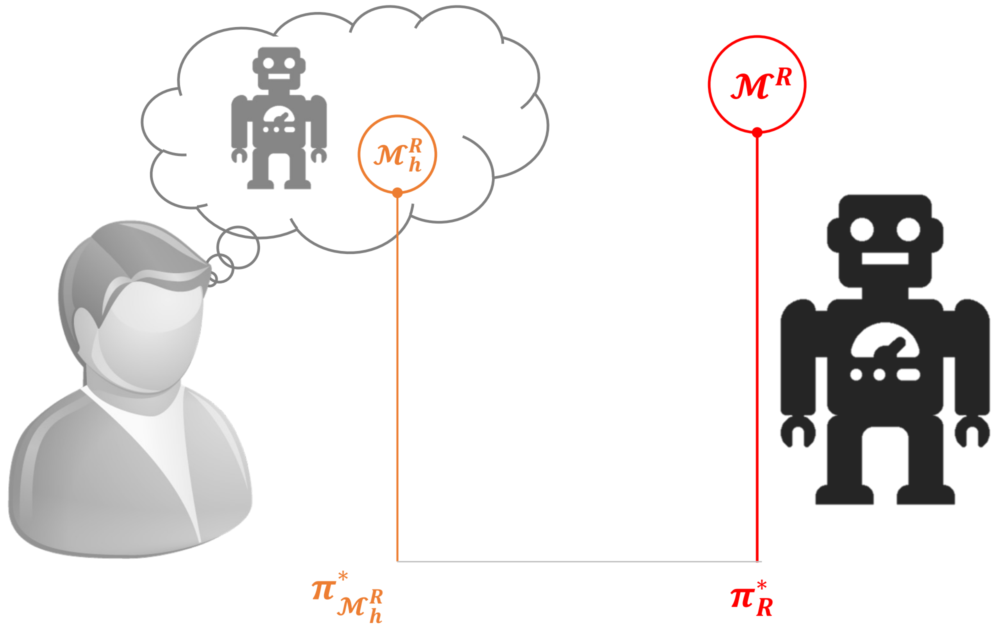


Model Reconciliation

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

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

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

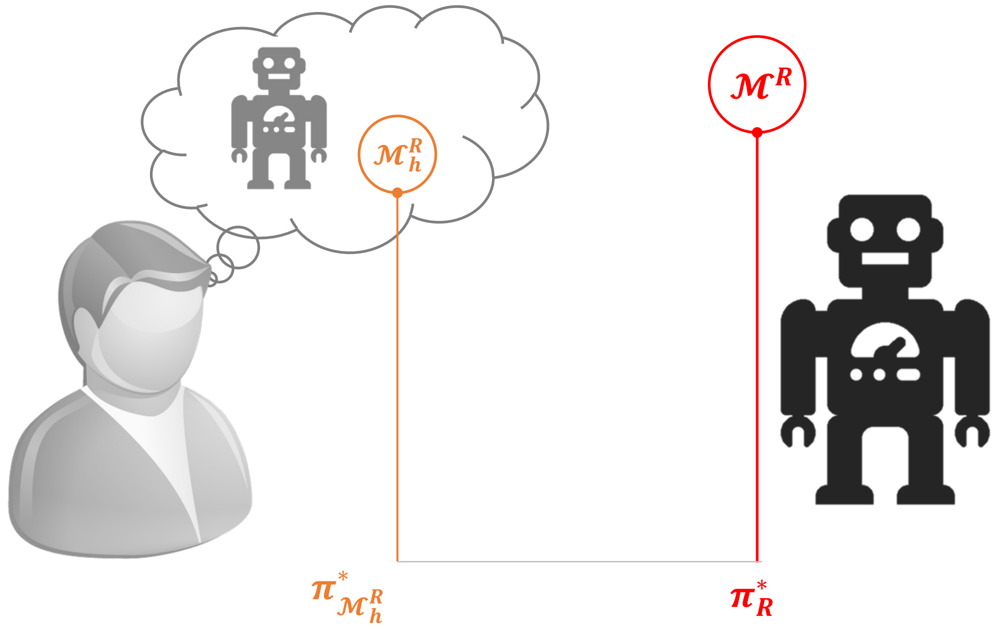


Why did you select π_R^*?

Model Reconciliation

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

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

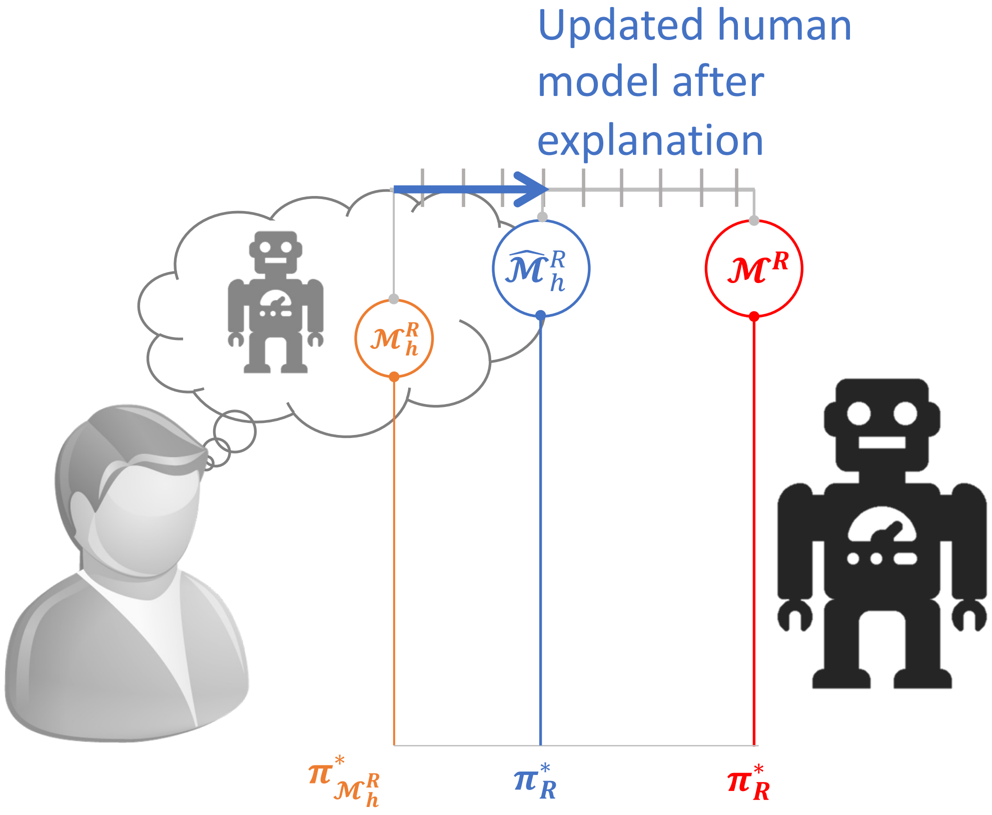


Why did you select π^*_R?

Model Reconciliation

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

Thus, our focus should be on identifying the minimal updates to be made to the human mental model so they can correctly evaluate the robot's plan.



Complexity Results

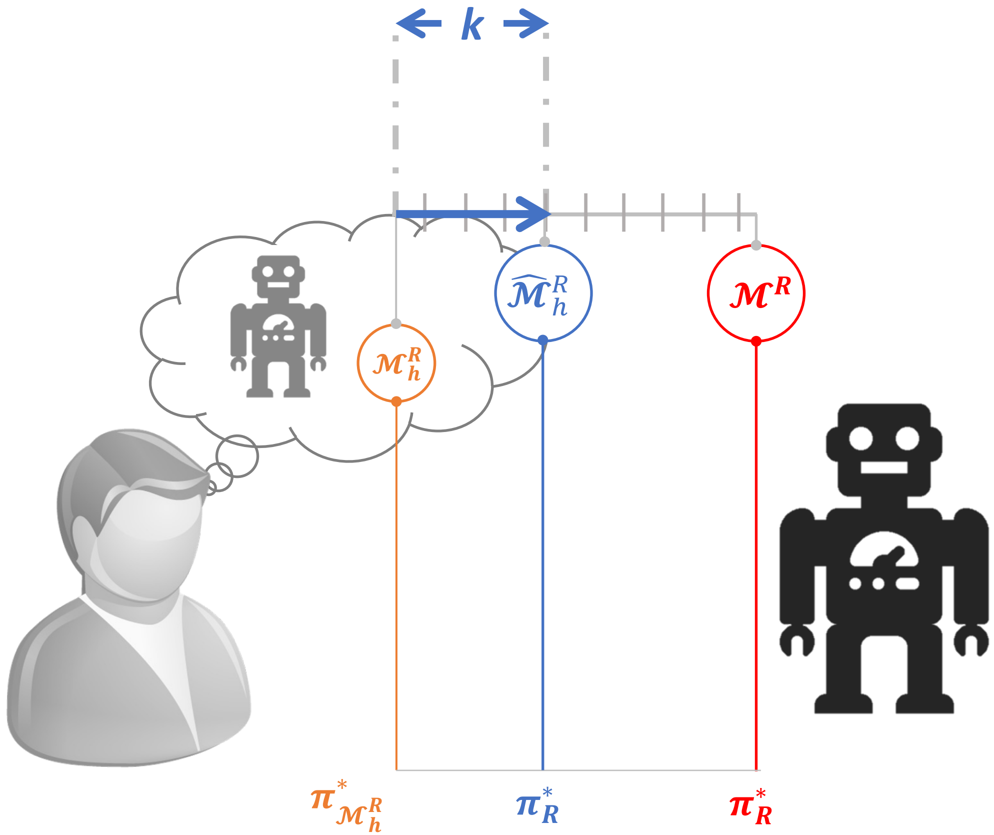
Complexity of the optimal model reconciliation explanation decision problem (*MRE*-k)

Does there exist a valid explanation of size k?

Where $\boxed{|\epsilon^{+}| + |\epsilon^{-}|} = k$

Model updates

Theorem 3. MRE- *k is Σ_2^p-Complete*



Σ_2^p Complexity

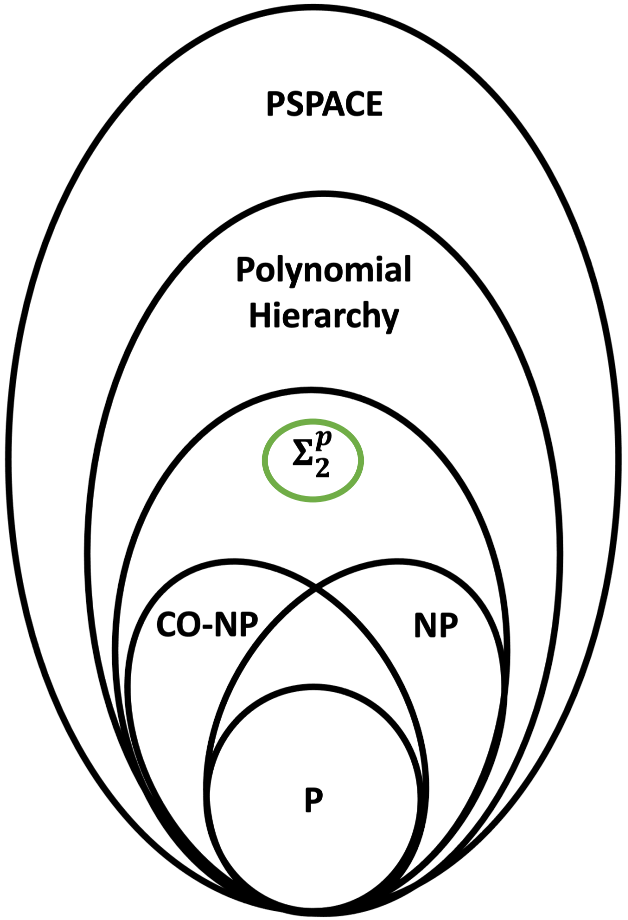
Part of the polynomial hierarchy

Placed between the PSPACE and NP (two classes that appear very commonly across planning problems)

Canonical problem: $QSAT_2$

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

A restricted class of quantified Boolean formulas



Proof Sketch for Main Complexity Results

***Theorem 3. MRE-**k is Σ_2^p-Complete*

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

MRE-k problem is compiled into a $QSAT_2$ problem

$$\langle M^R, M_h^R, \pi_R^* \rangle \longrightarrow \exists\, (X, Z) \forall\, Y\, (\,\phi_1(X) \wedge \neg\big(\phi_2(X, Y)\big) \wedge \phi_3(Z))$$

***Theorem 2. MRE-**k is Σ_2^p*-hard

MRE-k problem is compiled into a $QSAT_2$ problem

$$\exists X \forall Y\ \phi(X, Y) \rightarrow \exists X \neg(\exists Y\ \neg\phi(X, Y)) \longrightarrow \langle M^R, M_h^R, \pi_R^* \rangle$$

Existential quantifier encoded as possible model updates over initial states
Universal quantification 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

Take-Aways

- There exist a QBF compilation for model reconciliation explanation
  - Provided by the membership proof
  - Note that the compilation only leverages a subclass of the more general QBF problem

- Complexity is Σ_2^{p}-Complete