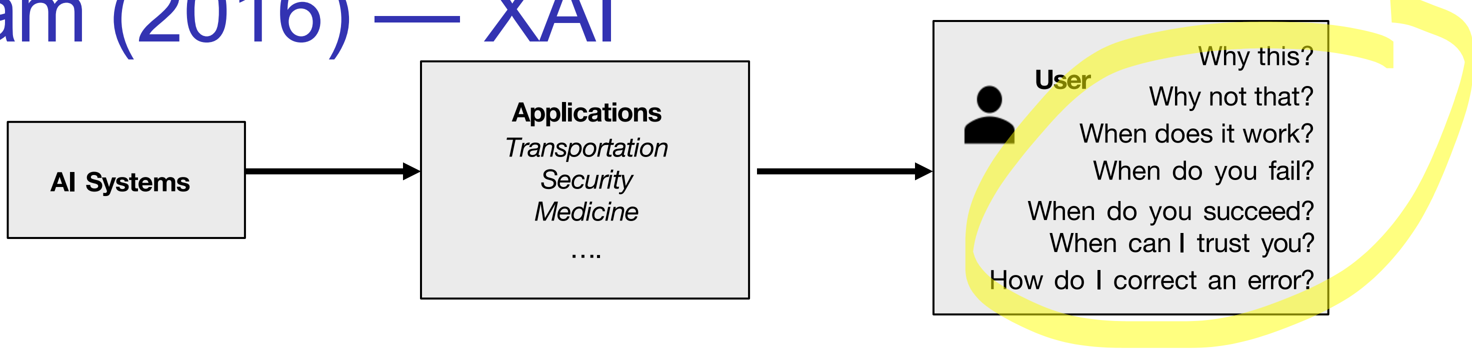
Model Reconciliation & Its Applications in Explainable AI

Tran Cao Son
New Mexico State University

Explainable AI

DARPA Program (2016) — XAI



Focus on machine learning systems — proposed theme:

- ❑ Producing explainable models (learning explanatory semantics: features, representations, structures, causal models, etc.) → [Producing explainable models, computing models]
- ❑ Designing explanation interface → [HCI]
- ❑ Understanding the psychological requirements for effective explanations → [Psychology]

https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

This Talk

- ❑ Explainable planning and model reconciliation
- ❑ Model reconciliation problem (MRP) under the lens of logic
- ❑ Solving MRP and applications of MRP

An explanation is a set of statements usually constructed to describe a set of facts which clarifies the causes, context, and consequences of those facts. This description may establish rules or laws and may clarify the existing rules or laws in relation to any objects, or phenomena examined.

Source: Introduction to Logic. Jess Drake (2018).

Explainable Planning

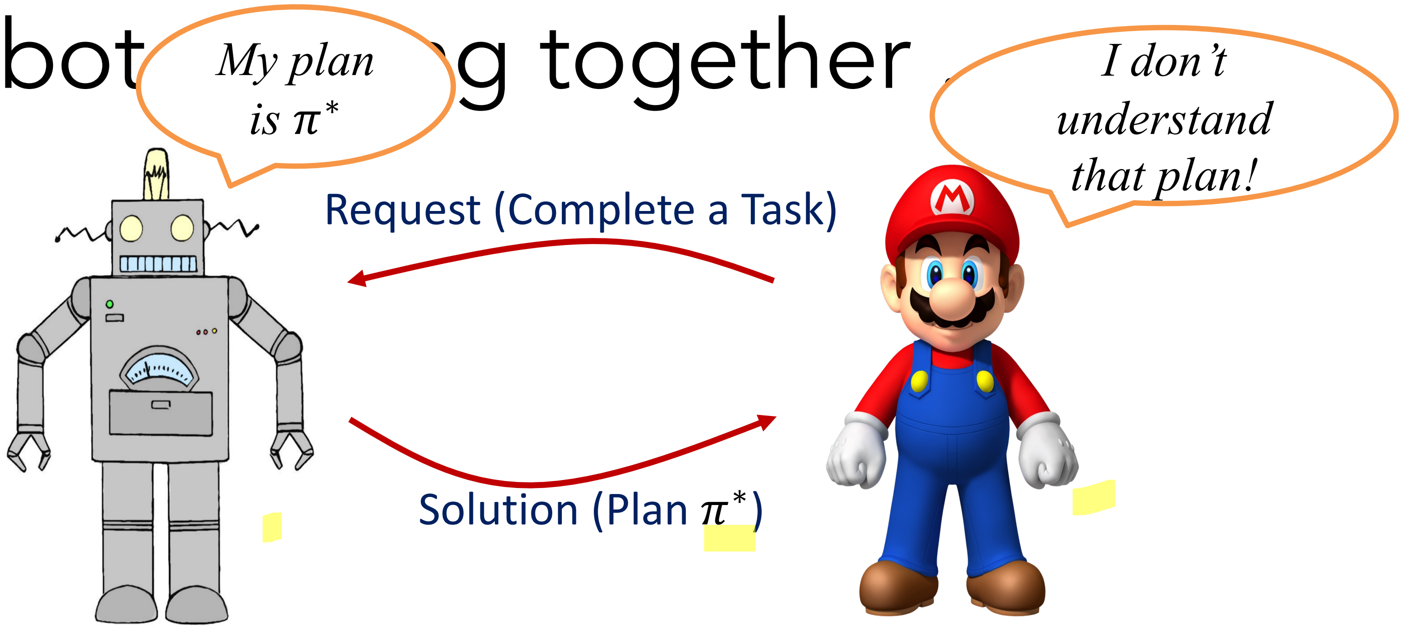
- Explainable AI in planning
- Relevant to xAI but also different from xAI
  - human and intelligent robots working together
  - white box vs. black box
- Different facets – who explains to whom
- Lot of interest and activities (e.g., xAIP workshop series: https://xaip.mybluemix.net)
- References: many in https://explainableplanning.com

Dimensions of Explainable Planning

- ❑ Planning model
  - ❑ classical planning
  - ❑ probabilistic planning
  - ❑ MDP/POMDP
  - ❑ …
- ❑ Knowledge of the explainer (robot)
  - ❑ both models ☞ one shot explanation
  - ❑ only its internal model ☞ dialog for explanation

Model Reconciliation in xAIP

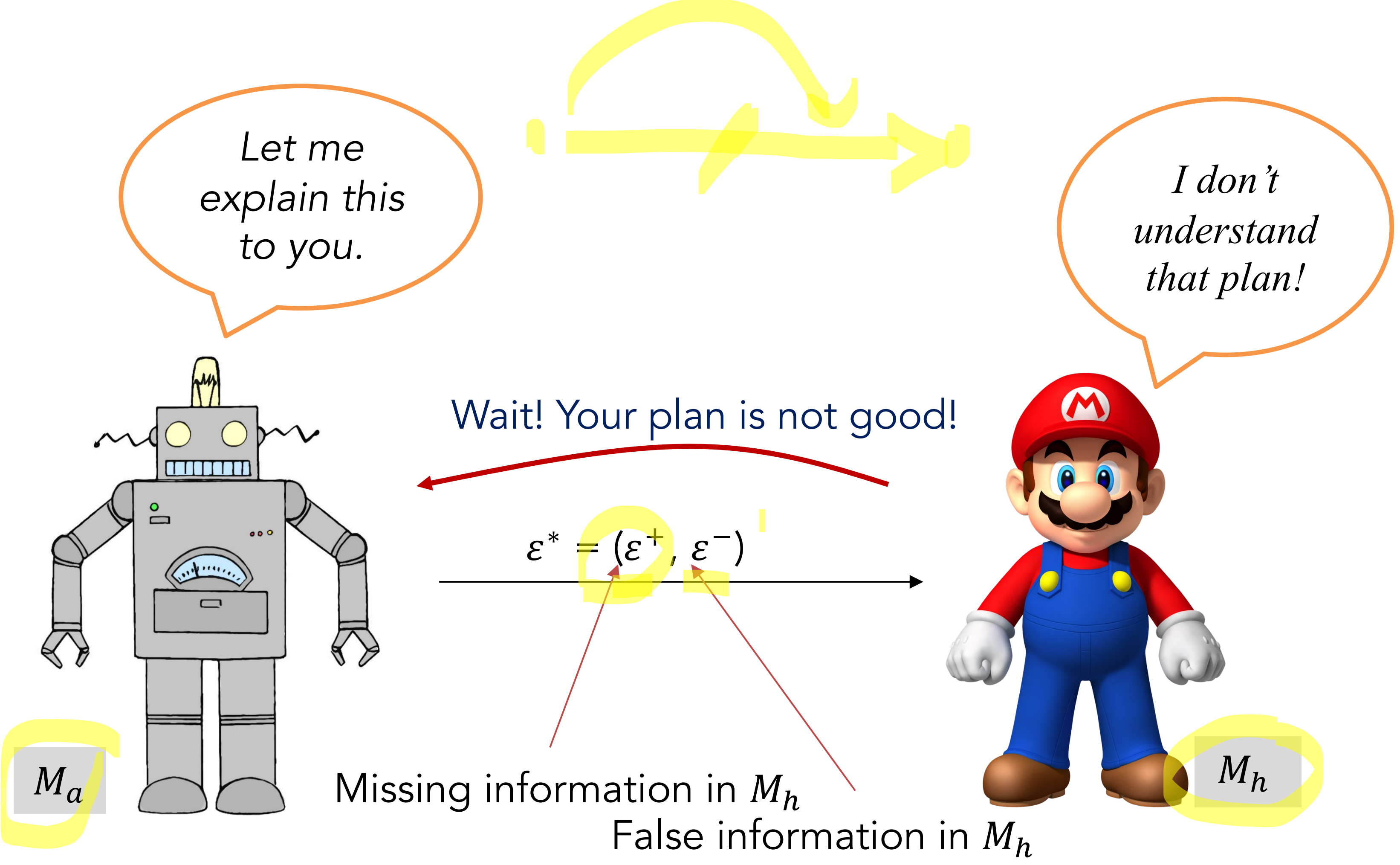
Human and robot … g together



- ❑ Human asks robot to accomplish a task and how the robot would complete it.
- ❑ Robot presents a plan.
- ❑ Human might have different plan π' *that is better* than the proposed plan.
- ❑ Explanation is needed!

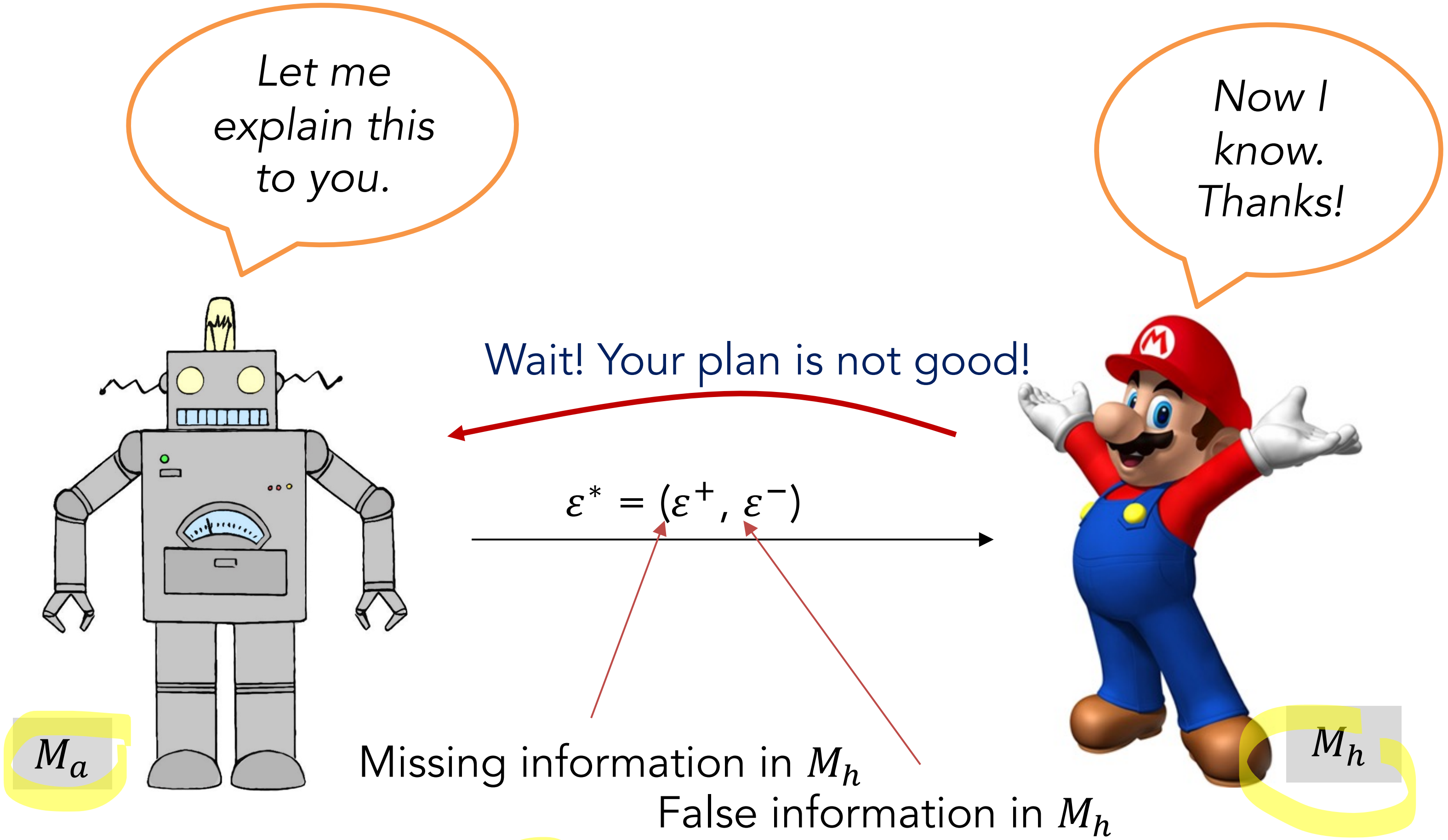
Model reconciliation as an approach to solving xAIP

Model Reconciliation in xAIP



M_x – planning model of x

Model Reconciliation in xAIP

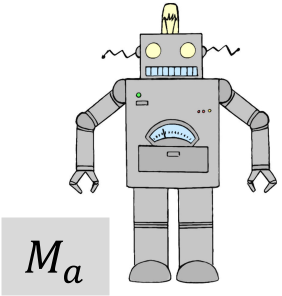


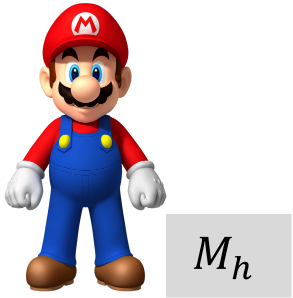
After reconciliation – $M_h \setminus \varepsilon^- \cup \varepsilon^+$ accepts the plan π^* of robot as valid

M_x – planning model of x

Formalization

- Two agents (robot and human) with two planning problems M_a and M_h
- Robot informs human about an optimal solution $\boldsymbol{\pi}^*$ for a goal G (written as $M_a \vDash \boldsymbol{\pi}^*$ and $M_h \nvDash \boldsymbol{\pi}^*$)
- Compute $\varepsilon = (\varepsilon^+, \varepsilon^-)$ such that $M_h \setminus \varepsilon^- \cup \varepsilon^+ \vDash \boldsymbol{\pi}^*$





$M \vDash \boldsymbol{\pi}$: $\boldsymbol{\pi}$ is a plan in the model M

Example – Block World Problem

Model of Robot

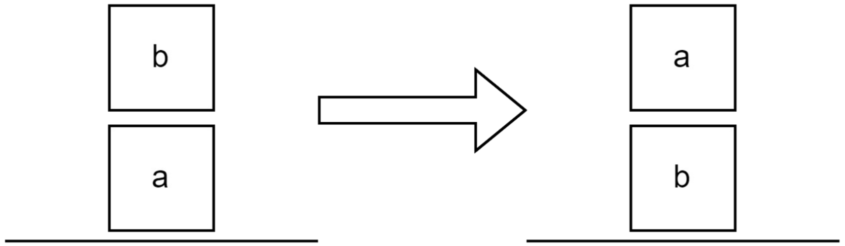
- stack(x,y)
  - Precondition: holding(x), clear(y)
  - Postcondition: handempty, on(x,y), ¬clear(y), clear(x)
- unstack(x,y)
  - Precondition: handempty, on(x,y) , clear(x)
  - Postcondition: holding(x), clear(y), ¬clear(x)
- …

Model of Human (missing precondition)

- stack(x,y)
  - Precondition: ~~holding(x)~~, clear(y)
  - Postcondition: handempty, on(x,y), ¬clear(y), clear(x)
- Other actions are as in the model of the robot.

Optimal Plan:

1. unstack(b,a)
2. putdown(b)
3. pickup(a)
4. stack(a,b)



initial state

goal state

Optimal Plan:

1. stack(a,b)

Explanation: missing holding(x) as a precondition of stack(x,y)

Research in Model Reconciliation in xAIP

- Computing minimal explanations
- One-shot: *the agent (robot) who computes an explanation knows both domain descriptions*
  - Specialized algorithms
  - Answer set programming-based algorithm
- Dialog (limited effort): *the agent (robot) who computes an explanation does not know the human domain description* – more realistic – later …

Computing Explanations for xAIP

Input: M_a, M_h, and $\boldsymbol{\pi}^*$ such that $M_a \vDash \boldsymbol{\pi}^*$ and $M_h \nvDash \boldsymbol{\pi}^*$

Output: An explanation $(\varepsilon^+, \varepsilon^-)$ for $(M_a, M_h, \boldsymbol{\pi}^*)$

repeat

non-deterministically select a potential $(\varepsilon^+, \varepsilon^-)$

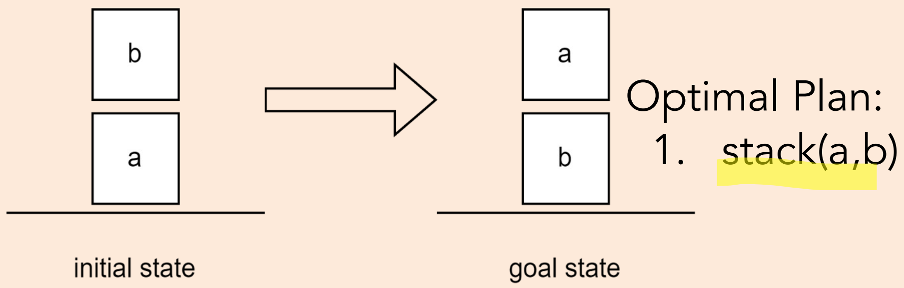
if $M_h \setminus \varepsilon^- \cup \varepsilon^+ \vDash \boldsymbol{\pi}^*$

then return $(\varepsilon^+, \varepsilon^-)$

until all possible explanations are considered

☚ how to select?

relevant to $\boldsymbol{\pi}^*$



Relevant: action stack(x,y) and initial state

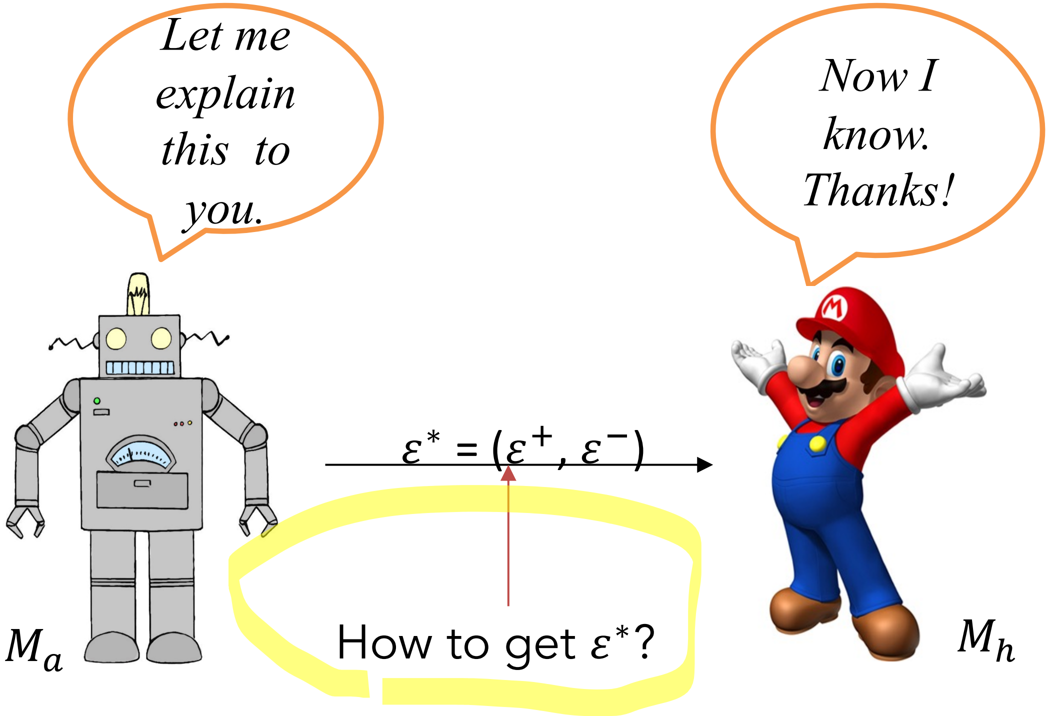
Note

- ❑ Only for one-shot explanation (knowledge of both M_a and M_h)
- ❑ ASP implementation comparable with state of the art (*Nguyen et al. - KR 2020*)
- ❑ SAT implementation (*Vasileiou et al. – JAIR 2022, AAAI 2021*)

From One-Shot Explanation to Dialog

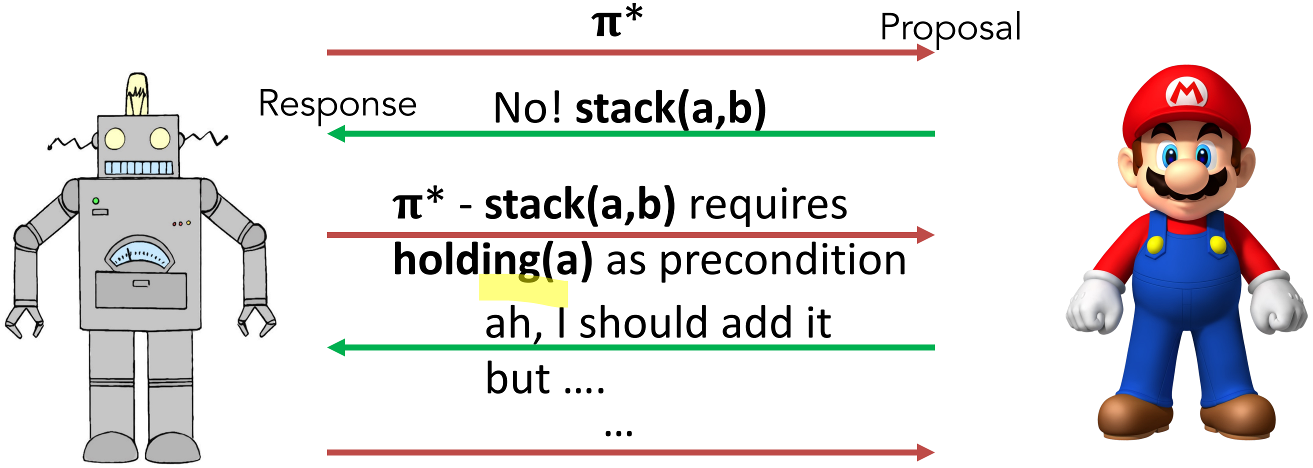
- **What if** *the agent (robot) who computes a solution does not know the domain description of the human*?

- Need to compute explanations by **talking** to each other, perhaps through multiple exchanges



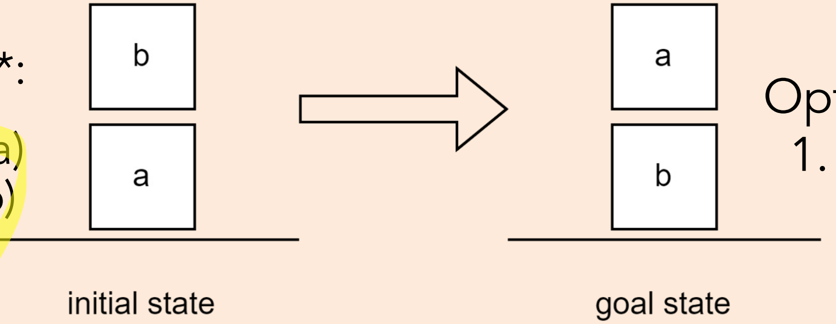
From One-Shot Explanation to Dialog

- Robot informs human of optimal plan **π*** and potential explanation **ε**
- Human reveals issues about **π*** and **ε**
  - the plan is not optimal
  - some action in **π*** cannot be executed in the updated model
  - Some goal cannot be achieved
  - …
- Robot receives the responses, identifies the issues, sends updated explanation



Optimal Plan **π***:

1. unstack(b,a)
2. putdown(b)
3. pickup(a)
4. stack(a,b)



Optimal Plan:

1. stack(a,b)

From One-Shot Explanation to Dialog

- Robot sends human optimal plan $\boldsymbol{\pi}^*$ and $\boldsymbol{\varepsilon}$ (initially, $\boldsymbol{\varepsilon}=\emptyset$)
- Human update M_h with $\boldsymbol{\varepsilon}$
  - the plan is not my solution, $\boldsymbol{\pi}'$ is better because
    - the plan $\boldsymbol{\pi}^*$ is not executable or does not achieve the goal
    - some action's precondition is relaxed or
    - some action's postcondition becomes true when it is not supposed to
  - some action cannot be executed
  - some goal cannot be achieved

  send the information to robot
- Robot identifying potential problems in human's response => create new $\boldsymbol{\varepsilon}$

Explanation and Proposal

- Given a model $M = (I, G, D)$ – as a set of atoms representing a planning problem
  - *add*(M) = {*add*(x) | x ∈ M}
  - *remove*(M) = {*delete*(x) | x is an atom representing an element of a planning model}
- $\boldsymbol{\varepsilon} \subseteq$ *add*(M) ∪ *remove*(M) is an **explanation** if there is no x such that *add*(x) ∈ $\boldsymbol{\varepsilon}$ and *remove*(x) ∈ $\boldsymbol{\varepsilon}$
- ($\boldsymbol{\pi}$, $\boldsymbol{\varepsilon}$) is a **proposal** w.r.t. M where $\boldsymbol{\pi}$ is an optimal solution for M and $\boldsymbol{\varepsilon}$ is an explanation

Response

($\boldsymbol{\pi}$, $\boldsymbol{\varepsilon}$) is a **proposal** w.r.t. M_a

$M_h \otimes \boldsymbol{\varepsilon} = M_h \setminus \{ x \mid remove(x) \in \boldsymbol{\varepsilon} \} \cup \{x \mid add(x) \in \boldsymbol{\varepsilon}\}$

A response for ($\boldsymbol{\pi}$, $\boldsymbol{\varepsilon}$) w.r.t. M_h

- *acceptable*: (T,T) if $M_h \otimes \boldsymbol{\varepsilon}$ has $\boldsymbol{\pi}$ as an optimal plan
- *non-optimal:* ($\boldsymbol{\pi}'$,T) if $M_h \otimes \boldsymbol{\varepsilon}$ has $\boldsymbol{\pi}'$ as a "better" plan than $\boldsymbol{\pi}$
- *redundant information:* ($\perp$, $\boldsymbol{\varepsilon}'$) where $\boldsymbol{\varepsilon}' = (add(M_h) \cap \boldsymbol{\varepsilon}) \cup \{remove(x) \mid x \notin M_h$ and $remove(x) \in \boldsymbol{\varepsilon}\}$
- *not executable:* (*, $\boldsymbol{\omega}$) where $\boldsymbol{\omega}$ encodes the information why $\boldsymbol{\pi}$ cannot be executed in $M_h \otimes \boldsymbol{\varepsilon}$

Dialog

Given (M_a, M_h, $\boldsymbol{\pi}^*$) a **dialogue** between robot and human is a sequence of rounds (x_i, y_i) where

- ❑ each $x_i = (\boldsymbol{\pi}^*, \boldsymbol{\varepsilon})$ is a proposal w.r.t. M_a
- ❑ each y_i is a response for x_i w.r.t. M_h

A dialog is **successful** if it is finite and the last response is an acceptable response (for both parties)

Note

- ❑ Clear separation of two parties (explainer and explainee)
- ❑ ASP implementation (*Ho & Son - ICLP 2022*)
- ❑ Argumentation based formalization (*Vasileiou et al. - upcoming*)

Model Reconciliation Problem (MRP)

Representation: knowledge bases of robot and human are represented by logical theories KB_a and KB_h in some logic L, respectively.

> $\models^c$ and $\models^s$ — credulous and skeptical entailment relationship between theories and formulae in L, respectively.

Exchange: question about the truth value of formulas over literals occurring in KB_a and KB_h.

Answer: suggested modification to the knowledge base of the questioner.

Model Reconciliation Problem (MRP)

Representation: knowledge bases of robot and human are represented by logical theories KB_a and KB_h in some logic L, respectively.

$\models^c$ and $\models^s$ — credulous and skeptical entailment relationship between theories and formulae in L, respectively.

Two types of questions (exchanges):

Entailment MRP (e-MRP): why does $KB_a \models^c \phi$ and $KB_h \models^s \neg\phi$?

Determining $\varepsilon^+ \subseteq KB_a$ and $\varepsilon^- \subseteq KB_h$ such that $(KB_h \setminus \varepsilon^-) \cup \varepsilon^+ \models^c \phi$.

non-Entailment MRP (n-MRP): why does $KB_a \models^s \phi$ and $KB_h \models^c \neg\phi$?

Determining $\varepsilon^+ \subseteq KB_a$ and $\varepsilon^- \subseteq KB_h$ such that $(KB_h \setminus \varepsilon^-) \cup \varepsilon^+ \models^s \phi$.

Compute a solution $(\varepsilon^+, \varepsilon^-)$ of a MRP (KB_a, KB_h, ϕ).

Connection to xAIP

$KB_{a/h}$ = collection of formulae encoding the transition function for reasoning about effects of actions
plus the description of the initial state
plus the description of the goal

Well-known SAT or ASP encoding

$KB_a \models^c \phi$ (goal) ~ extracted plan for ϕ exists in SAT/ASP – length is a parameter of KB

- ❑ α an optimal plan for the robot and not the human
  - ❑ e-MRP: $KB_a \models^c \phi$ *but* $KB_h \not\models^s \phi$

Characterizing Explanations

Given (KB_a, KB_h, ϕ) and an explanation $(\varepsilon^+, \varepsilon^-)$

- $(\varepsilon^+, \varepsilon^-)$ is *optimal* if there exists no explanation (λ^+, λ^-) such that $\lambda^+ \cup \lambda^- \subset \varepsilon^+ \cup \varepsilon^-$. (always exists)
- $(\varepsilon^+, \varepsilon^-)$ is π-*restrictive* for $\pi \subseteq KB_a$ if $\varepsilon^+ \subseteq \pi$. (existence depends on π)
  - *minimally-restrictive* if there exists no explanation (λ^+, λ^-) such that $\lambda^+ \subset \varepsilon^+$. (always exists)
- $(\varepsilon^+, \varepsilon^-)$ is π-*preserving* for $\pi \subseteq KB_h$ if $\pi \cap \varepsilon^- = \emptyset$; (existence depends on π)
  - *maximally-preserving* if there exists no explanation (λ^+, λ^-) such that $\lambda^- \subset \varepsilon^-$. (always exists)

Cost-Based Characterization

Cost function $C: KB_a \cup KB_h \rightarrow R^{\geq 0}$ (elements in KB_a and KB_h are representation unit).

$$C(\varepsilon^+, \varepsilon^-) =_{def} \sum_{r \in \varepsilon^+ \cup \varepsilon^-} C(r)$$

$(\varepsilon^+, \varepsilon^-)$ **cost optimal** w.r.t. C if $C(\varepsilon^+, \varepsilon^-)$ is minimal among all explanation.

Given C of (KB_a, KB_h, ϕ)

- *uniform* if C(r) = c (constant, > 0) for r ∈ $KB_a \cup KB_h$

cost-optimal explanation w.r.t. C is optimal

- *agent-biased* if C(r) = c for r ∈ KB_a and C(r) = 0 for r ∈ KB_h.

cost-optimal explanation w.r.t. C is minimally-restrictive

- *human-biased* if C(r) = 0 for r ∈ KB_a and $C(r)$ = c for r ∈ KB_h.

cost-optimal explanation w.r.t. C is maximally-preserving

Computing Explanation: What is Needed?

- Given: Logic L, formula $\phi = \phi^+ \wedge \phi^-$, KB_a and KB_h
  ($KB_a \models^c \phi^+$ and $KB_a \models^s \phi^-$) and ($KB_h \not\models^c \phi^+$ or $KB_h \not\models^s \phi^-$)
- Compute $(\varepsilon^+, \varepsilon^-)$ so that $KB_h \setminus \varepsilon^- \cup \varepsilon^+ \models^c \phi^+$ and $KB_h \setminus \varepsilon^- \cup \varepsilon^+ \models^s \phi^-$
- How to
  - $\models^c$ and $\models^s$? (logic dependent)
  - $KB_h \setminus \varepsilon^- \cup \varepsilon^+$ (updating a logical theory, logic and structural dependent)

Propositional logic: $\models^c$ (SAT) and $\models^s$ (UNSAT)
ASP: $\models^c$ (one answer set, brave) and $\models^s$ (all answer sets, skeptical)
Argumentation Framework: $\models^c$ (credulous) and $\models^s$ (skeptical)

Computing Explanation: What is Needed?

- Given: Logic L, formula $\phi = \phi^+ \wedge \phi^-$, KB_a and KB_h
  ($KB_a \models^c \phi^+$ and $KB_a \models^s \phi^-$) and ($KB_h \not\models^c \phi^+$ or $KB_h \not\models^s \phi^-$)
- Compute $(\varepsilon^+, \varepsilon^-)$ so that $KB_h \setminus \varepsilon^- \cup \varepsilon^+ \models^c \phi^+$ and $KB_h \setminus \varepsilon^- \cup \varepsilon^+ \models^s \phi^-$
- How to
  - $\models^c$ and $\models^s$? (logic dependent)
  - $KB_h \setminus \varepsilon^- \cup \varepsilon^+$ (updating a logical theory, logic and structural dependent)

Propositional logic: ?
ASP: ?
Argumentation Framework: ?

→ Updating operator KB $\otimes$ $(\varepsilon+, \varepsilon-)$

Computing Explanation

Input: Logic L, formula $\phi = \phi^+ \wedge \phi^-$, KB_a and KB_h

$(KB_a \models^c \phi^+$ and $KB_a \models^s \phi^-)$ and $(KB_h \not\models^c \phi^+$ or $KB_h \not\models^s \phi^-)$

Output: A solution for (KB_a, KB_h, ϕ)

repeat

non-deterministically select a potential $(\varepsilon^+, \varepsilon^-)$

if $KB_h \otimes \varepsilon^- \cup \varepsilon^+ \models^c \phi^+$ and $KB_h \otimes \varepsilon^- \cup \varepsilon^+ \models^s \phi^-$

then return $(\varepsilon^+, \varepsilon^-)$

until all possible explanations are considered

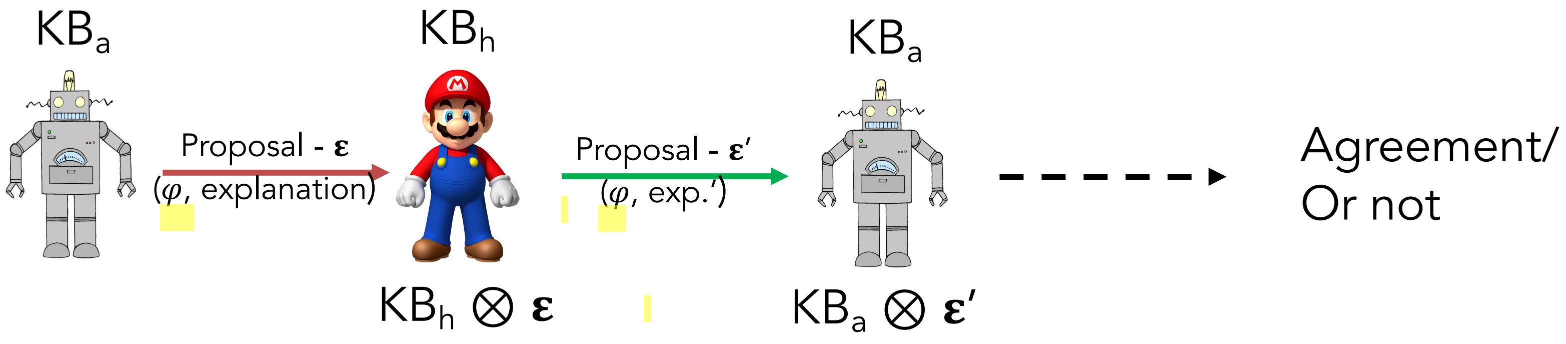
- Only for one-shot situation (knowledge of both KB_a and KB_h)

Dialog for Model Reconciliation

Explanation – sub-theory

Proposal – (φ, explanation)

Dialog – sequence of proposal exchanges



Is model reconciliation the same as dialog/negotiation?

Towards a full ASP Implementation of MRP

- Propositional logic program **Π**, rule of the form

  $r: \quad a_0 \leftarrow a_1, \ldots, a_m, \textbf{not } a_{m+1}, \ldots, \textbf{not } a_n$

  $head(r)$: a_0, $pos(r)$: $\{a_1, \ldots, a_m\}$, $neg(r)$: $\{a_{m+1}, \ldots, a_n\}$
- Π^S: **reduction** of **Π** with respect to a set of atoms S.
- T_Π: **immediate consequence operator** for positive programs.
- $S = lfp(T_{\Pi^S})$: answer set.
- S is a **justification** for q with respect to an answer set I if $S \subseteq \Pi$ such that $head(r) \in I$ and $I \models body(r)$ for $r \in S$ and $q \in lfp(TSI)$.

What is available?

- $\Pi \models^c a$: a in some answer set of **Π**
- $\Pi \models^s a$: a in all answer sets of **Π**
- **ε** – rules and facts

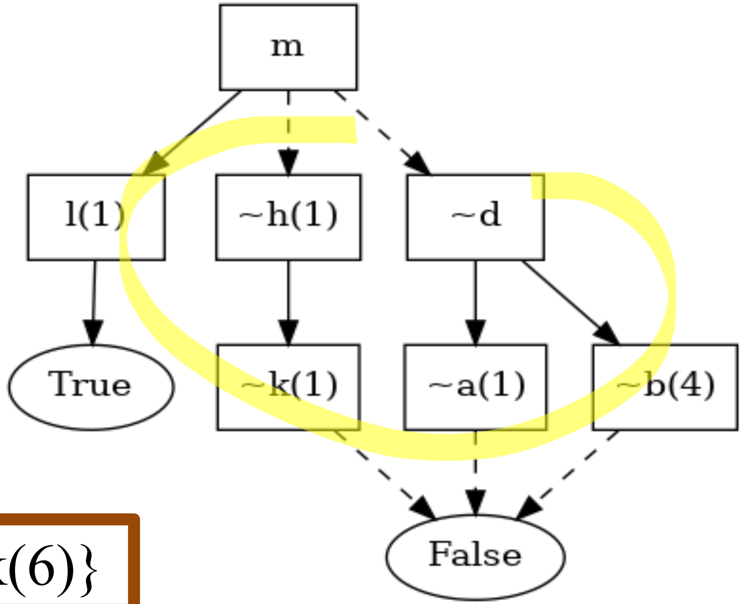
What is missing?

- **Task 1**: given $\Pi \models^c a$ and answer set S, why is $a \in S$ (or $a \notin S$)?
- **Task 2**: $\Pi \otimes \varepsilon$: how to update?

Computing Explanation

Given an ASP program **Π**, an atom a, an answer set S.
Compute an explanation for a being in (or not in) S.

```
(r1) m :- l(X), not d, not h(X).
(r2) d :- b(X), a(X).
(r3) h(X) :- k(X), p.
(r4) b(1).
(r5) a(4).
(r6) l(1).
(r7) k(6).
```



A = {l(1), m, a(4), b(1), k(6)}

Several system ICLP21/LPNMR22
Newest system considers almost all syntactic constructors of modern ASP solver

(Alviano et al. ICLP 2023)

Computing $\Pi \otimes \boldsymbol{\varepsilon}$? Intuition

$\pi = \{a \leftarrow\}$

$(b, \boldsymbol{\varepsilon}=(\{b \leftarrow a\}, \emptyset))$

$\pi \otimes \boldsymbol{\varepsilon} = \{a \leftarrow; b \leftarrow a\}$

$\pi = \{a \leftarrow\}$

$(b, \boldsymbol{\varepsilon}=(\{b \leftarrow \text{not } a\}, \emptyset))$

$\pi \otimes \boldsymbol{\varepsilon} = \{b \leftarrow \text{not } a\}$

$\pi = \{c \leftarrow \text{not } c\}$

$(b, \boldsymbol{\varepsilon}=(\{b \leftarrow \text{not } a\}, \{c \leftarrow \text{not } c\})$

$\pi \otimes \boldsymbol{\varepsilon} = \{b \leftarrow \text{not } a\}$

What to keep and what to remove?

A Proposal for Update Operator

For a set I of atoms and a set of rule $\boldsymbol{\varepsilon}^+$ (w.r.t some program π_a), the *residual* of π with respect to $\boldsymbol{\varepsilon}^+$ and I, $\otimes(\pi, \boldsymbol{\varepsilon}^+, I) \subseteq \pi \setminus \boldsymbol{\varepsilon}^+$, such that for each rule $r \in \otimes(\pi, \boldsymbol{\varepsilon}^+, I)$:

- ❑ $head(r) \in I$ and $neg(r) \cap I = \emptyset$; or
- ❑ $neg(r) \cap heads(\boldsymbol{\varepsilon}^+) \neq \emptyset$; or
- ❑ $pos(r) \setminus I \neq \emptyset$.

Define: $\boldsymbol{\varepsilon}^-[\boldsymbol{\varepsilon}^+, I, \pi] =_{def} \pi \setminus \otimes(\pi, \boldsymbol{\varepsilon}^+, I)$

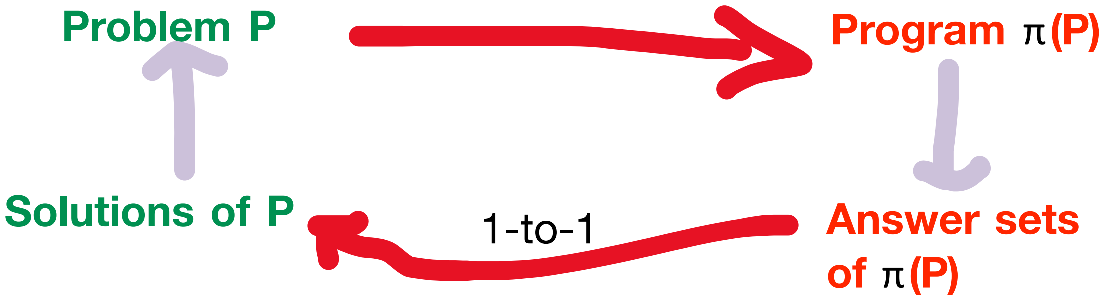
If $\boldsymbol{\varepsilon}^+$ is a justification for q then $(\boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-[\boldsymbol{\varepsilon}^+, I, \pi])$ is an explanation for q w.r.t. π (i.e., $\pi \setminus \boldsymbol{\varepsilon}^-[\boldsymbol{\varepsilon}^+, I, \pi] \cup \boldsymbol{\varepsilon}^+ \models q$)

☞ ε^- can be determined given ε^+, I, π

ASP implementation (*JELIA 2021*)

MRP and Its Applications in ASP

- P encoded as a set of facts FA
  - planning: the domain, the initial state, and the goal are encoded as a set of facts
  - scheduling: the set of tasks and constraints
  - graph coloring: the set of nodes and edges



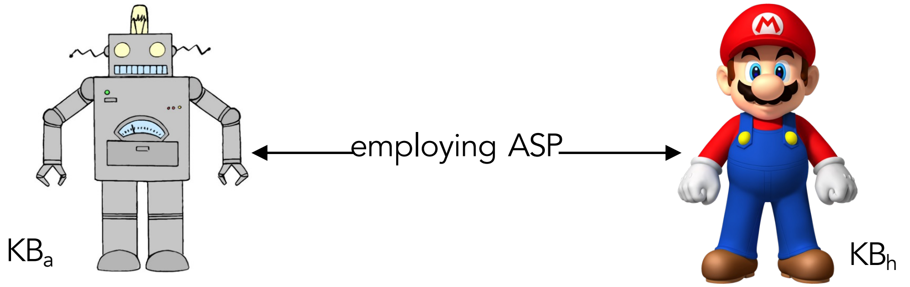
- $\pi(P) = FA \cup R(FA)$
  - planning: domain independent rules for reasoning about actions' executability and effects, etc.
  - scheduling: rules for assigning tasks to resource and time and checking satisfiability of solutions
  - graph coloring: rules for assigning colors to nodes and checking satisfiability of solutions

- R(FA): obtained from grounding domain independent rules using ground terms in FA

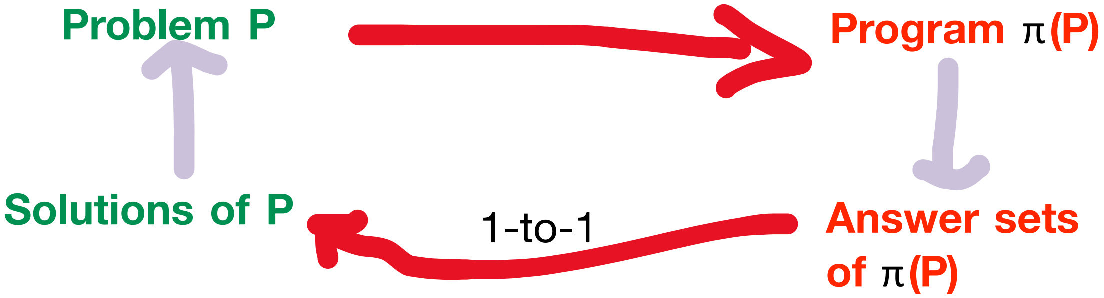
MRP and Its Applications in ASP

When robot and human employ ASP for problem solving and

- ❑ they share the set of domain independent rules
- ❑ MRP between KB_a ($\pi(P_a)$) and KB_h ($\pi(P_h)$) reduces to MRP between P_h and P_a



  - ❑ Planning
  - ❑ Scheduling
  - ❑ Graph coloring
  - ❑ …



General algorithm for ASP is applicable but scalability is an issue.

Related Work

- Earlier problems such as logic program update, diagnosis, or explanation in abductive logic programs could be viewed as instances of MRP – one shot explanation
- Multiagent diagnosis
- Negotiation
- Dialog

Logic programming update

$P \otimes \phi$: ϕ is new information and needs to be integrated to P

ϕ must be true in $P \otimes \phi$, maintaining as much information from P as possible. focus: how to construct $P \otimes \phi$

Update obeys belief revision principles (*causal rejection principle* or *program transformation*)

- individual rules might change
- $P \otimes \phi$ might be inconsistent
- $P \otimes \phi$ might be a set of programs

Instance of MRP: (ϕ, P, ϕ): $P = \pi_h$, $\pi_a = \phi$

Explanation in Abductive Logic Programs

Given (P, A) and a query q.

Identify a pair of hypotheses (E, F) that explains q:

- $(P \setminus F) \cup E$ is consistent
- $E \subseteq A \setminus P$ and $F \subseteq A \cap P$
- $(P \setminus F) \cup E \models q$

Instance of MRP: (A, P, q): $P = \pi_h$, $A = \pi_h \cup \pi_a$, $\pi_a \models q$.

Diagnosis

Given a theory KB, a set of observations OBS, a set of predefined atoms AB(C) such that KB ∪ OBS ∪ AB(C) is inconsistent.

Identify a set D ⊆ AB(C) such that KB ∪ OBS ∪ (AB(C) \ D) ∪ ¬D is consistent.

Instance of MRP: (¬AB(C), KB ∪ OBS ∪ AB(C), True)

Summary & Outlook

- Model reconciliation for explainable planning
  - One-shot explanation
  - Dialog for explanation
- Logic based formalization of MRP
  - Characterizations of explanation of MRP
  - Method for computing explanation of MLP using ASP
- Future work
  - Multiagent diagnosis
  - Negotiation
  - Dialog

Acknowledgements

- Collaborators:
  - Washington U: William Yeoh, Stylianos Vasileiou
  - St Joseph U: Marcello Balduccini
  - U of Calabria: Mario Alviano
- NMSU Students
  - Van Nguyen, Ly Ly Trieu, Dung Ho
- NSF support

Thank you!