Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy

Authors: Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, Subbarao Kambhampati

Speaker: Abetare Shabani

University of Luxembourg

December 2020

Outline

- Introduction
- Related work
- Properties of explanations
- Algorithm presentation
- Evaluation of performance of algorithms
- Conclusion & Future work
- Citations

Motivation & Goal

- Different model used by AI and different from human
- AI systems are mostly called to explain their plans and behaviors
- The authors believe that explanations are best explained in light of model differences

- Goal: Explanation as Model Reconciliation problem

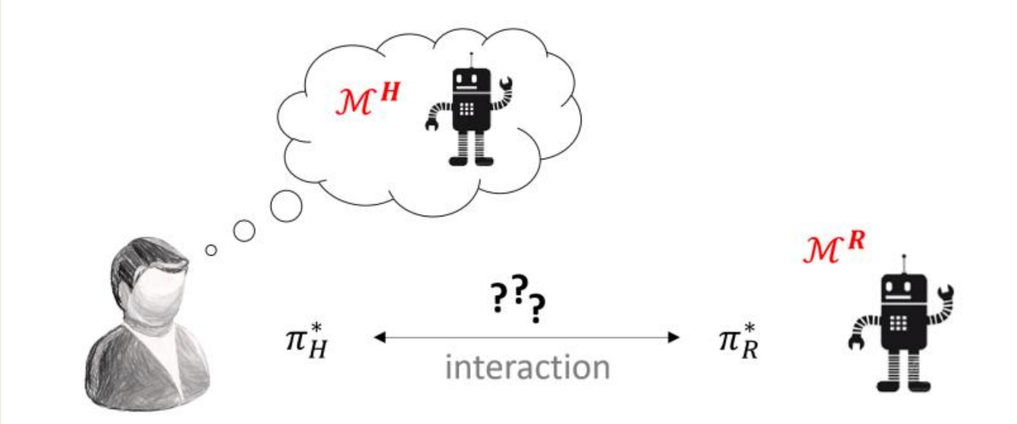
HUMAN IN THE LOOP

Why and How?

https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0

Introduction

- Explanation to humans-in-the-loop
- Earlier work
  - Planner explaining decision on respect of its own model
  - What issues can encounter?
- Explanations should be robot’s attempt to change human’s model to correspond to its plan



Contribution

- Model explanation as Model Reconciliation Problem
- Robot’s optimal plan
- New model-search algorithms
- Explanation generation system

Multi-model Setting Scenario

```
(: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) ) )
```

Fetch Robot

https://fetchrobotics.com/

Related Work

- The work is supported by psychology studies
  - Lombrozo, 2006,2012.

- Optimal plan – valid and better than other alternatives
- Different from other model change algorithms
- Most of the work done involved humans entering the land of planners

Classical Planning Problem

- Planner's plan comprehensible to humans
- M = $\langle D, I, G \rangle$
  - D = $\langle F, A \rangle$ – domain
- Solution – $\pi = \langle a_1, a_2, \ldots, a_{\mathrm{n}} \rangle$
- π* known as the cheapest plan
- Optimal plan not always is optimal in M_{H}

Multi Model Planning Setting

- Tuplet of $\langle M_{\mathrm{H}}, M_{\mathrm{R}} \rangle$
- Two approaches
  1. Change its own behavior in order to be explicable to the human
  2. Bring the human's model closer to its own

Model Reconciliation Problem

- Tuple $\langle \pi^* , \langle M_{\mathrm{H}}, M_{\mathrm{R}} \rangle \rangle$
- Mapping function Γ: $M \rightarrow s$
- Model change actions can make only one change at a time
- Solution - edit functions $\{\lambda_{\mathrm{i}}\}$ that can transform $\mathrm{M}_1 \rightarrow \mathrm{M}_2$

Multi Model Explanations

- Plan is more optimal in the updated model than in original one
- The update of the model can be negotiated by humans
- Each solution for this problem requires these :
  - Completeness
  - Conciseness
  - Monotonicity
  - Computability

Plan Patch Explanation

- Incomplete
- Limitation: ignores model differences, contains information that does not need to be revealed
- Solution : provide the entire model difference to the human

Model Patch Explanation

- Easy to compute
- Limitation: far from being concise due to large size
- Goal: minimize the size

Minimally Complete Explanation

- Shortest complete explanation
- Fetch Robot the smallest example of MCE
- Human can compute the optimal plan given a planning problem.

Model-space search for MCE

- Equal importance to all model corrections
- Proposition 1: selection strategy of successor nodes to speed up search
- Proposition 2: feasibility of the plan in the modified planning problem is a necessary but not a sufficient condition for a valid explanation

Algorithm 1 Search for Minimally Complete Explanations

procedure MCE-SEARCH
Input: MRP $\langle \pi^*, \langle \mathcal{M}^R, \mathcal{M}^H \rangle \rangle$
Output: Explanation $\mathcal{E}^{MCE}$
Procedure:
fringe $\leftarrow$ `Priority_Queue()`
c_list $\leftarrow \{\}$ ▷ Closed list
$\pi_R^* \leftarrow \pi^*$ ▷ Optimal plan being explained
$\pi_H \leftarrow \pi$ such that $C(\pi, \mathcal{M}^H) = C^*_{\mathcal{M}^H}$ ▷ Plan expected by human
fringe.push($\langle \mathcal{M}^H, \{\} \rangle$, priority $= 0$)
while True **do**
$\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \leftarrow$ fringe.pop($\widehat{\mathcal{M}}$)
if $C(\pi_R^*, \widehat{\mathcal{M}}) = C^*_{\widehat{\mathcal{M}}}$ **then return** $\mathcal{E}$ ▷ Return $\mathcal{E}$ if π_R^* optimal in $\widehat{\mathcal{M}}$
else
c_list $\leftarrow$ c_list $\cup\ \widehat{\mathcal{M}}$
for $f \in \Gamma(\widehat{\mathcal{M}}) \setminus \Gamma(\mathcal{M}^R)$ **do** ▷ Models that satisfy condition 1
$\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{\}, \{f\} \rangle$ ▷ Removes f from $\widehat{\mathcal{M}}$
if $\delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$ c_list **then**
fringe.push($\langle \delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle$, $c + 1$)
for $f \in \Gamma(\mathcal{M}^R) \setminus \Gamma(\widehat{\mathcal{M}})$ **do** ▷ Models that satisfy condition 2
$\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\}, \{\} \rangle$ ▷ Adds f to $\widehat{\mathcal{M}}$
if $\delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$ c_list **then**
fringe.push($\langle \delta_{\mathcal{M}^H, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle$, $c + 1$)

procedure PRIORITY_QUEUE.POP($\hat{\mathcal{M}}$)
candidates $\leftarrow \{\langle \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c^* \rangle \mid c^* = \arg\min_c \langle \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \rangle\}$
pruned_list $\leftarrow \{\}$
$\pi_H \leftarrow \pi$ such that $C(\pi, \hat{\mathcal{M}}) = C^*_{\hat{\mathcal{M}}}$
for $\langle \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \rangle \in$ candidates **do**
if $\exists a \in \pi_R^* \cup \pi_H$ such that $\tau^{-1}(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\hat{\mathcal{M}})) \in \{c_a\} \cup pre(a) \cup eff^+(a) \cup eff^-(a)$ **then** ▷ Candidates relevant to π_R^* or π_H
pruned_list $\leftarrow$ pruned_list $\cup\ \langle \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \rangle$
if pruned_list $= \phi$ **then** $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \sim Unif$(candidate_list)
else $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \sim Unif$(pruned_list)

Minimally Monotonic Explanation

- Preserves completeness and monotonicity
- Proposition 3: MME solution is equal to the differences between M and M_{R}
- Proposition 4: MMEs are not unique to an MRP problem.
- Proposition 5: MCE may not be a subset of an MME

Model Search for MME

- Search over the entire model space
- Goal: find the largest set of model changes for which the explicability criterion becomes invalid for the first time

Algorithm 2 Search for Minimally Monotonic Explanations

procedure MME-SEARCH
Input: MRP $\langle \pi^*, \langle \mathcal{M}^R, \mathcal{M}^H \rangle \rangle$
Output: Explanation $\mathcal{E}^{MME}$
Procedure:
  $\mathcal{E}^{MME} \leftarrow \{\}$
  fringe $\leftarrow$ `Priority_Queue()`
  c_list $\leftarrow \{\}$ ▷ Closed list
  h_list $\leftarrow \{\}$ ▷ List of incorrect model changes
  fringe.push($\langle \mathcal{M}^R, \{\} \rangle$, priority $= 0$)
  while fringe is not empty **do**
    $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \leftarrow$ fringe.pop($\widehat{\mathcal{M}}$)
    if $C(\pi^*, \widehat{\mathcal{M}}) > C^*_{\widehat{\mathcal{M}}}$ **then**
      h_list $\leftarrow$ h_list $\cup$ $(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R))$ ▷ Updating h_list
    else
      c_list $\leftarrow$ c_list $\cup \widehat{\mathcal{M}}$
      for $f \in \Gamma(\widehat{\mathcal{M}}) \setminus \Gamma(\mathcal{M}^H)$ **do** ▷ Models that satisfy condition 1
        $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{\}, \{f\} \rangle$ ▷ Removes f from $\widehat{\mathcal{M}}$
        if $\delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$ c_list
          and $\nexists S$ s.t. $(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)) \supseteq S \in$ h_list **then** ▷ Prop 3
          fringe.push($\langle \delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle, c + 1$)
          $\mathcal{E}^{MME} \leftarrow \max_{|\cdot|}\{\mathcal{E}^{MME}, \mathcal{E}\}$
      for $f \in \Gamma(\mathcal{M}^H) \setminus \Gamma(\widehat{\mathcal{M}})$ **do** ▷ Models that satisfy condition 2
        $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\}, \{\} \rangle$ ▷ Adds f from $\widehat{\mathcal{M}}$
        if $\delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$ c_list
          and $\nexists S$ s.t. $(\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)) \supseteq S \in$ h_list **then** ▷ Prop 3
          fringe.push($\langle \delta_{\mathcal{M}^R, \mathcal{M}^H}(\Gamma(\widehat{\mathcal{M}}), \lambda), \mathcal{E} \cup \lambda \rangle, c + 1$)
          $\mathcal{E}^{MME} \leftarrow \max_{|\cdot|}\{\mathcal{E}^{MME}, \mathcal{E}\}$
  $\mathcal{E}^{MME} \leftarrow (\Gamma(\widehat{\mathcal{M}}) \Delta \Gamma(\mathcal{M}^R)) \setminus \mathcal{E}^{MME}$
  return $\mathcal{E}^{MME}$

Evaluation

- Explanation generation system
  - For planning: Fast-Downward
  - Plan validation: VAL
  - Parsing: Pyperplan
- The experiment was run on a 12 core system
- Planning domains: BlocksWorld, Logistics and Rover

- No completeness guarantee but better computability of an explanation.
- Replace the equality test:
  1. π_{R}^{*} is valid in the new hypothesis model
  2. The new plan has become better or at least π_{H} is diproved.
  3. Each action contributes at least one causal link to π* in M.
- Proposition 6: Criterion 3 is necessary for optimality of π* in M

Domain Name	Problem	MPE (ground truth)		PPE		MME (exact)		MCE (exact w/o heuristic)		MCE (exact with heuristic)		MCE (approximate)	
		size	time	size	time	size	time	size	time	size	time	size	time
BlocksWorld	1	10	n/a	5	n/a	3	1100.8	2	34.7	2	18.9	2	19.8
	2			8		4	585.9	3	178.4	3	126.6	3	118.8
	3			4		5	305.3	2	34.7	2	11.7	2	11.7
	4			7		5	308.6	3	168.3	3	73.3	3	73.0
Rover	1	10	n/a	10	n/a	2	2093.2	2	111.3	2	100.9	2	101.0
	2			10		2	2018.4	2	108.6	2	101.7	2	102.7
	3			10		2	2102.4	2	104.4	2	104.9	2	102.5
	4			9		1	3801.3	1	13.5	1	12.8	1	12.5
Logistics	1	5	n/a	5	n/a	4	13.7	4	73.2	4	73.5	4	63.6
	2			5		4	13.5	4	73.5	4	71.4	4	63.3
	3			5		5	8.6	5	97.9	5	100.4	3	36.4
	4			5		5	8.7	5	99.2	5	95.4	3	36.4

$\lvert \mathcal{M}^R \Delta \mathcal{M}^H \rvert$	problem-1	problem-2	problem-3	problem-4
3	2.2	18.2	4.7	18.5
5	6.0	109.4	15.4	110.2
7	7.3	600.1	23.3	606.8
10	48.4	6849.9	264.2	6803.6

BlocksWorld	problem-1	problem-2	problem-3	problem-4
Number of nodes expanded for MME (out of 1024)	128	64	32	32

RESULTS

Conclusion & Future work

- Explanations in this multi-model setting become a process of identifying and reconciling the relevant differences between the models

- Future work
  - human’s models that are of different form than the robot’s, to allow effective learning of the human’s models

- Limitations
  - Explanations must be compatible with the planner’s model
  - The Robots acknowledge human model to come up with optimal plan
  - The level of abstraction in the Human model.

Citations

- 127 citations
- From which 37 in 2020
- The latest work
  - The Emerging Landscape of Explainable Automated Planning & Decision Making

Thank you for your attention!