Explicability versus Explanations in Human-Aware Planning*

Robotics Track†

Tathagata Chakrabroti, Sarath Sreedharan, Subbarao Kambhampati
Arizona State University
Tempe, AZ 85281 USA
[tchakra2,ssreedh3,rao]@asu.edu

ABSTRACT

Human aware planning requires an agent to be aware of the mental model of the human in the loop during its decision process. This can involve generating plans that are explicable to the human as well as the ability to provide explanations when such plans cannot be generated. In this paper, we bring these two concepts together and show how an agent can account for both these needs and achieve a trade-off during the plan generation process itself by means of a model-space search method MEGA*. This provides a revised perspective of what it means for an AI agent to be "human-aware" by bringing together recent works on explicable planning and plans explanations under the umbrella of a single plan generation process. We illustrate these concepts using a robot involved in a typical search and reconnaissance task with an external supervisor.

KEYWORDS

Human-Aware Planning; Plan Explicability; Plan Explanations; Model Reconciliation; Argumentation

ACM Reference Format:
Tathagata Chakrabroti, Sarath Sreedharan, Subbarao Kambhampati. 2018. Explicability versus Explanations in Human-Aware Planning. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 HUMAN-AWARE PLANNING

A Classical Planning Problem [8, 10] is a tuple $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}, \mathcal{G} \rangle$ with domain $\mathcal{D} = \langle F, A \rangle$ – where F is a finite set of fluents that define a state $s \subseteq F$, and A is a finite set of actions – and initial and goal states $\mathcal{I}, \mathcal{G} \subseteq F$. Action $a \in A$ is a tuple $\langle c_a, pre(a), eff^{\pm}(a) \rangle$ where c_a is the cost, and $pre(a), eff^{\pm}(a) \subseteq F$ are the preconditions and add/delete effects, i.e. $\delta_{\mathcal{M}}(s, a) \models \bot$ *if* $s \not\models pre(a)$; *else* $\delta_{\mathcal{M}}(s, a) \models s \cup eff^{+}(a) \setminus eff^{-}(a)$ where $\delta_{\mathcal{M}}(\cdot)$ is the transition function.

The solution to the planning problem is a sequence of actions or a (satisficing) *plan* $\pi = \langle a_1, a_2, \ldots, a_n \rangle$ such that $\delta_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$. The cost of a plan π is given by $C(\pi, \mathcal{M}) = \sum_{a \in \pi} c_a$ if $\delta_{\mathcal{M}}(\mathcal{I}, \pi) \models \mathcal{G}$; ∞ otherwise. The cheapest plan $\pi^* = \arg\min_{\pi} C(\pi, \mathcal{M})$ is the (cost) optimal plan, whose cost is denoted by $C^*_{\mathcal{M}}$.

A Human-Aware Planning (HAP) Problem is given by the tuple $\Psi = \langle \mathcal{M}^R, \mathcal{M}^R_h \rangle$ where $\mathcal{M}^R = \langle D^R, \mathcal{I}^R, \mathcal{G}^R \rangle$ is the planner's model of a task, while $\mathcal{M}^R_h = \langle D^R_h, \mathcal{I}^R_h, \mathcal{G}^R_h \rangle$ is the human's understanding of the same (i.e. the human mental model).

Thus, a human-aware agent incorporates the *human mental model* [3] in addition to the its own model in its deliberative process in order to anticipate how its plans are *perceived* from the point of view of the human in the loop. For example, an immediate consequence of differences between the planner's model and the human mental model is that optimal plans produced by the planner are no longer optimal when evaluated in the human mental model and thus may be considered *inexplicable* by the human.

Explicable Planning – An "explicable" solution to an HAP is a plan π such that (1) it is executable (but may no longer be optimal) in the planner's model but is (2) "closer" to the optimal (and hence, expected) plan in the human mental model –

(1) $\delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \models \mathcal{G}^R$; and
(2) $C(\pi, \mathcal{M}^R_h) \approx C^*_{\mathcal{M}^R_h}$.

"Closeness" or distance to the expected plan is modeled here in terms of cost optimality, but in general this can be any metric such as plan similarity. In existing literature [7, 12, 13] this has been achieved by modifying the search process so that the heuristic that guides the search is driven by the human mental model.

Plan Explanations – The other approach would be to compute optimal (and possibly inexplicable) plans and provide an explanation in terms of the differences with the human mental model that causes this inexplicability. This is referred to as the *model reconciliation process* [5, 11] which provides an (1) explanation or model update $\mathcal{E}$ such that the (2) optimal plan is (3) also optimal (and hence, explained) in the updated human mental model –

(1) $\widehat{\mathcal{M}^R_h} \longleftarrow \mathcal{M}^R_h + \mathcal{E}$; and
(2) $C(\pi, \mathcal{M}^R) = C^*_{\mathcal{M}^R}$;
(3) $C(\pi, \widehat{\mathcal{M}^R_h}) = C^*_{\widehat{\mathcal{M}^R_h}}$.

2 EXPLICABILITY *VERSUS* EXPLANATIONS

Indeed, these two processes of explanations and explicability are intrinsically related in an agent's deliberative process. For example, an agent can generate a explicable plan to the best of its ability or it can provide explanations whenever required, or it can even opt for a combination of both – e.g. if the expected human plan is too costly in the planner's model (e.g. the human might not be aware of some safety constraints) or the cost of communication overhead for

*The first two authors contributed equally.
† The full version of the paper is available at https://arxiv.org/abs/1708.00543.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

explanations is too high (e.g. limited communication bandwidth). In the following discussion, we try to attain the sweet spot in this explanations versus explicability tradeoff.

From the perspective of design of autonomy, the explicability versus explanations trade-off has two interesting implications – (1) The agent can now not only explain but also *plan* in the multi-model setting with the trade-off between compromise on its optimality and possible explanations in mind; and (2) By incorporating the explanation generation process into an agent's decision making process itself, we mimic an argumentation process that is known to be a crucial function of the reasoning capabilities of humans [9]. Indeed, general argumentation frameworks for resolving disputes over plans have been explored before [2, 6]. Our work can be seen as the specific case where the argumentation process is over a set of constraints that prove the correctness and quality of plans by considering the cost of the argument specifically as it relates to the trade-off in plan quality and the cost of explaining that plan. *This is the first of its kind algorithm that can achieve this.*

A Balanced Solution – The result of a trade off in the relative cost of explicability and explanations is a plan π and an explanation $\mathcal{E}$ such that (1) π is executable in the agent model, and with the explanation (2) in the form of model updates it is (3) optimal in the updated human model while (4) the cost (length) of the explanations and the cost of deviation from optimality in its own model to be explicable to the human is traded off according to a constant α –

(1) $\delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \models \mathcal{G}^R$;
(2) $\widehat{\mathcal{M}}_h^R \longleftarrow \mathcal{M}_h^R + \mathcal{E}$;
(3) $C(\pi, \widehat{\mathcal{M}}_h^R) = C^*_{\widehat{\mathcal{M}}_h^R}$; and
(4) $\pi = \arg\min_\pi \ |\mathcal{E}| \ + \ \alpha \times \mid C(\pi, \mathcal{M}^R) - C^*_{\mathcal{M}^R} \mid$.

With higher values of α the agent will prefer plans that require more explanation, while with lower α it will be more explicable.

2.1 The MEGA* Algorithm

We employ a *model space* A^* search (Algorithm 1) to compute the expected plan and explanations for a given value of α. Similar to [5, 11] we define a state representation over planning problems with a mapping function $\Gamma : \mathcal{M} \mapsto \mathcal{F}$ which represents a planning problem by transforming every condition in it into a predicate. The set Λ of actions contains unit model change actions $\lambda : \mathcal{F} \rightarrow \mathcal{F}$ which make a single change to a domain at a time. We start by initializing the min node tuple ($\mathcal{N}$) with the human mental model and an empty explanation. For each new possible model, we test if the objective value of the new node is smaller than the current min node. We stop the search once we identify a model that is capable of producing a plan that is also optimal in the robot's own model. This is different from [5], where we were just trying to identify the first model where a *given plan* is optimal.

Property. MEGA* yields the smallest possible explanation for a given HAP. This is beyond what is offered by [5], which only computes the smallest explanation *given* a plan.

Property. $\alpha = \mid \mathcal{M}^R \Delta \mathcal{M}_h^R \mid$ yields the most optimal plan along with the minimal explanation in a given HAP. $\alpha = 0$ yields the most explicable plan. This is distinct from just computing the optimal plan in the human mental model, since such a plan may not be executable so that some explanations are required even in the worst case. This is a welcome additions to the explicability only view of plan generation introduced in [7, 12, 13], where the human model only guides plan generation but provides no insight into how to make the remainder of the model reconciliation possible.

Algorithm 1 MEGA*

procedure MEGA*-SEARCH
Input: HAP $\Psi = \langle \mathcal{M}^R, \mathcal{M}_h^R \rangle$, α
Output: Plan π and Explanation $\mathcal{E}$
Procedure:
  fringe ← Priority_Queue()
  c_list ← {} ▷ Closed list
  $\mathcal{N}_{min} \leftarrow \langle \mathcal{M}_h^R, \{\} \rangle$ ▷ Node with minimum objective value
  $\pi^*_R \leftarrow \pi^*$ ▷ Optimal plan being explained
  $\pi_h^R \leftarrow \pi$ s.t. $C(\pi, \mathcal{M}_h^R) = C^*_{\mathcal{M}_h^R}$ ▷ Plan expected by human
  fringe.push($\langle \mathcal{M}_h^R, \{\}\rangle$, priority = 0)
  while True **do**
    $\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle, c \leftarrow$ fringe.pop($\widehat{\mathcal{M}}$)
    if OBJ_VAL($\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle$) ≤ OBJ_VAL($\mathcal{N}_{min}$) **then**
      $\mathcal{N}_{min} \leftarrow \langle \widehat{\mathcal{M}}, \mathcal{E} \rangle$ ▷ Update min node
    if $C(\pi^*_{\widehat{\mathcal{M}}}, \mathcal{M}^R) = C^*_{\mathcal{M}^R}$ **then**
      $\langle \mathcal{M}_{min}, \mathcal{E}_{min} \rangle \leftarrow \mathcal{N}_{min}$
      return $\langle \pi_{\mathcal{M}_{min}}, \mathcal{E}_{min} \rangle$ ▷ If $\pi^*_{\widehat{\mathcal{M}}}$ is optimal in $\mathcal{M}^R$
    else
      c_list ← c_list ∪ $\widehat{\mathcal{M}}$
      for $f \in \Gamma(\widehat{\mathcal{M}}) \setminus \Gamma(\mathcal{M}^R)$ **do** ▷ Models that satisfy Condition 1 [5]
        $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{\}, \{f\} \rangle$ ▷ Removes f from $\widehat{\mathcal{M}}$
        if $\delta_{\mathcal{M}_h^R, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$ c_list **then**
          fringe.push($\langle \delta_{\mathcal{M}_h^R, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda), \ \mathcal{E} \cup \lambda \rangle, \ c + 1$)
      for $f \in \Gamma(\mathcal{M}^R) \setminus \Gamma(\widehat{\mathcal{M}})$ **do** ▷ Models that satisfy Condition 2 [5]
        $\lambda \leftarrow \langle 1, \{\widehat{\mathcal{M}}\}, \{f\}, \{\} \rangle$ ▷ Adds f to $\widehat{\mathcal{M}}$
        if $\delta_{\mathcal{M}_h^R, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda) \notin$ c_list **then**
          fringe.push($\langle \delta_{\mathcal{M}_h^R, \mathcal{M}^R}(\Gamma(\widehat{\mathcal{M}}), \lambda), \ \mathcal{E} \cup \lambda \rangle, \ c + 1$)
  procedure OBJ_VAL($\langle \widehat{\mathcal{M}}, \mathcal{E} \rangle$)
    return $|\mathcal{E}| \ + \ \alpha \times \mid C(\pi^*_{\widehat{\mathcal{M}}}, \mathcal{M}^R) - C^*_{\mathcal{M}^R} \mid$

Demonstration – We provide a demonstration of MEGA* in a typical [1] search and reconnaissance (USAR) scenario where a remote robot is assigned tasks by an external (human) commander. *A video can be viewed at https://youtu.be/Yzp4FU6Vn0M*. We show how, for low α, MEGA* chooses a plan that requires the least amount of explanation, i.e. the most explicable plan. This requires only a single initial state explanation to make the plan seem optimal but the robot must perform a costly rubble removal action to clear a path that the human expects to be accessible. The robot switches to the optimal plan for higher values of α along with a longer explanation updating the human of the evolved state of the world.

User Study – We also conducted an extensive human-factors study [4] to evaluate how these explanations are received by humans in the loop. The salient findings of the study as it relates to the explicability-explanations trade-off are available in the full report†.

Acknowledgments. This research is supported in part by the AFOSR grant FA9550-18-1-0067, the ONR grants N00014161-2892, N00014-13-1-0176, N00014- 13-1-0519, N00014-15-1-2027, and the NASA grant NNX17AD06G. Chakraborti is also supported by the IBM Ph.D. Fellowship 2016-18.

REFERENCES

[1] Cade Earl Bartlett. 2015. Communication between Teammates in Urban Search and Rescue. *Masters Thesis* (2015). Arizona State University.

[2] Alexandros Belesiotis, Michael Rovatsos, and Iyad Rahwan. 2010. Agreeing on plans through iterated disputes. In *AAMAS*. 765–772.

[3] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. 2017. AI Challenges in Human-Robot Cognitive Teaming. *arXiv preprint arXiv:1707.04775* (2017).

[4] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati. 2018. Plan Explanations as Model Reconciliation – An Empirical Study. *ArXiv e-prints* (Feb. 2018). arXiv:cs.AI/1802.01013

[5] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*.

[6] Chukwuemeka D Emele, Timothy J Norman, and Simon Parsons. 2011. Argumentation strategies for plan resourcing. In *AAMAS*.

[7] Anagha Kulkarni, Tathagata Chakraborti, Yantian Zha, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. 2016. Explicable Robot Planning as Minimizing Distance from Expected Behavior. *CoRR* abs/1611.05497 (2016).

[8] Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. 1998. PDDL-the planning domain definition language. (1998).

[9] Hugo Mercier and Dan Sperber. 2010. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences* (2010).

[10] Stuart Russell and Peter Norvig. 2003. *Artificial intelligence: a modern approach*. Prentice Hall.

[11] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2018. Handling Model Uncertainty and Multiplicity in Explanations as Model Reconciliation. In *ICAPS*.

[12] Yu Zhang, Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2016. Plan Explicability for Robot Task Planning. In *RSS Workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics*.

[13] Yu Zhang, Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2017. Plan Explicability and Predictability for Robot Task Planning. In *ICRA*.