

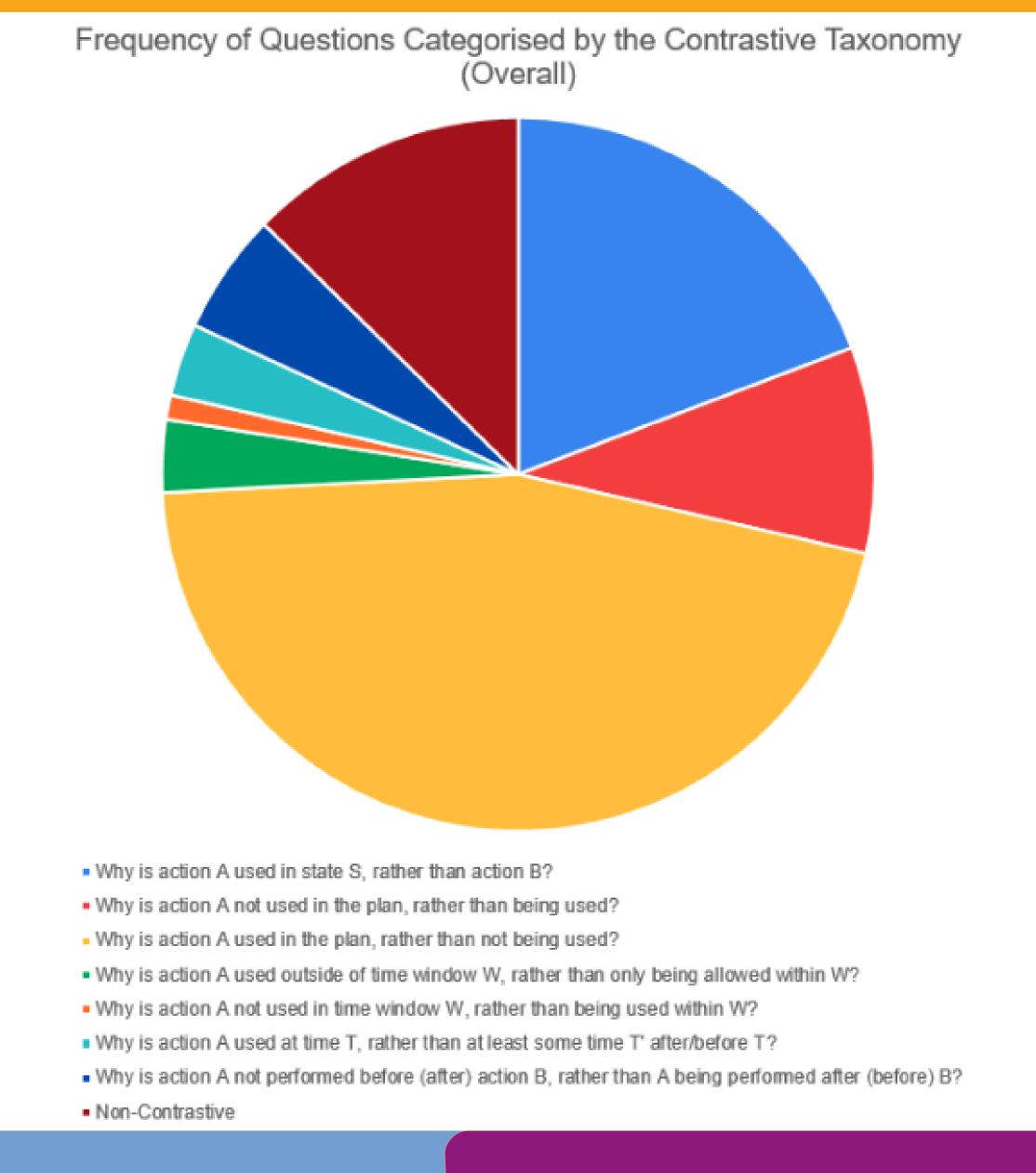
Contrastive Explanations of Plans Through Model Restrictions Journal of Artificial Intelligence Research, Volume 72, 2021, 533-621

Benjamin Krarup, Senka Krivic, Daniele Magazzeni, Derek Long, Michael Cashmore, David E. Smith King's College London, UK, benjamin.krarup@kcl.ac.uk

In automated planning, the need for explanations arises when there is a mismatch between a proposed plan and the user's expectation. We frame Explainable AI Planning as an iterative plan exploration process, in which the user asks a succession of contrastive questions that lead to the generation and solution of hypothetical planning problems that are restrictions of the original problem. The object of the exploration is for the user to understand the constraints that govern the original plan and, ultimately, to arrive at a satisfactory plan. We present the results of a user study that demonstrates that when users ask questions about plans, those questions are usually contrastive, i.e. "why A rather than B?". We use the data from this study to construct a taxonomy of user questions that often arise during plan exploration. Our approach to iterative plan exploration is a process of successive model restriction. Each contrastive user question imposes a set of constraints on the planning problem, leading to the construction of a new hypothetical planning problem as a restriction of the original. Solving this restricted problem results in a plan that can be compared with the original plan, admitting a contrastive explanation. We formally define model-based compilations in PDDL2.1 for each type of constraint derived from a contrastive user question in the taxonomy, and empirically evaluate the compilations in terms of computational complexity. The compilations were implemented as part of an explanation framework supporting iterative model restriction. We demonstrate its benefits in a second user study.

Framework Constraint Question XAIPService HModel Model HModel Generation Iterative Process HPlan Planner Non-Contrastive Synthesis Validate Plan HPlan XAIP Service Contrastive Explanation HPlan **XAIP Service** VAL report for HPlan Remained actions Added actions Rescheduled actions Removed actions Save HPlan Do you want to know: Why is action A not involved in the Original Plan: pp HPlan obtained by adding action (unload_pallet tom p2 sh1) 0.000: (goto_waypoint tom sh5 sh6) [3.000] Why is action A involved in the 0.000: (load_pallet jerry p1 sh3) [2.000] 2.000: (goto waypoint jerry sh3 sh4) [5.000] 000: (goto_waypoint tom sh5 sh6) [3.000] 0.000: (goto_waypoint tom sh5 sh6) [3.000] Why action A rather than action I 000: (load_pallet jerry p1 sh3) [2.000] 3.001: (set_shelf tom sh6) [1.000] 0.000: (load_pallet jerry p1 sh3) [2.000] 000: (goto_waypoint jerry sh3 sh4) [5.000] 000: (goto_waypoint jerry sh3 sh4) [5.000] 4.001: (goto_waypoint tom sh6 sh1) [4.000] Why is action A used before/afte 1: (set shelf tom sh6) [1.000] 001: (set shelf tom sh6) [1.000] 7.001: (goto_waypoint jerry sh4 sh5) [1.000] 001: (goto_waypoint tom sh6 sh1) [4.000] 001: (goto waypoint tom sh6 sh1) [4.000] 8.001: (set_shelf tom sh1) [1.000] Why action A at this time? .001: (goto_waypoint jerry sh4 sh5) [1.000] 01: (goto_waypoint jerry sh4 sh5) [1.000] 8.002: (goto_waypoint jerry sh5 sh6) [3.000] Why is action A used outside of 001: (set_shelf tom sh1) [1.000] 001: (set_shelf tom sh1) [1.000] 9.001: (goto waypoint tom sh1 sh2) [4.000] 001: (goto_waypoint tom sh1 sh6) [4.000] 11.002: (unload_pallet jerry p1 sh6) [1.500] Why is action A used at time t 3.001: (load_pallet tom p2 sh6) [2.000] 12.503: (load_pallet jerry p2 sh6) [2.000] (unload_pallet jerry p1 sh6) [1.500] 5.001: (goto waypoint tom sh6 sh1) [4.000] 14.503: (goto_waypoint jerry sh6 sh1) [4.000] 18.503: (unload_pallet jerry p2 sh1) [1.500] (load_pallet jerry p2 sh6) [2.000] 001: (unload_pallet tom p2 sh1) [1.500] (goto_waypoint jerry sh6 sh1) [4.000] 0.002: (goto_waypoint jerry sh5 sh6) [3.000] Cost: 23.502 Cost: 20.003000 Validation: plan valid Validation: plan valid Is this your question? Why is action (unload_pallet tom p2 sh1) not involved in the plan?

Why Contrastive Explanations?



Example Compilation

$$\Pi' = \langle \langle Ps', Vs, As', arity' \rangle, \langle Os, I', G', W' \rangle \rangle$$

where:

- $Ps' = Ps \cup \{can_do_a, not_done_a\}$
- $\bullet \ As' = \{o_a, o_{\neg a}\} \cup As \setminus \{o\}$
- $arity'(x) = arity(x), \forall x \in arity$
- $arity'(can_do_a) = arity'(not_done_a) = arity'(o_a) = arity'(o_{\neg a}) = arity(o)$
- $I' = I \cup \{ground(not_done_a, \chi)\}$
- $G' = G \cup \{ground(not_done_a, \chi)\}$
- $W' = W \cup \{\langle lb, ub, ground(can_do_a, \chi) \rangle\}$

where the new operators $o_{\neg a}$ and o_a extend o with the delete effect not_done_a and the precondition can_do_a , respectively. i.e:

$$Eff_{\vdash}^{-}(o_{\neg a}) = Eff_{\vdash}^{-}(o) \cup \{not_done_a\}$$

 $Pre_{\vdash}(o_a) = Pre_{\vdash}(o) \cup \{can_do_a\}$

Iterative Process

