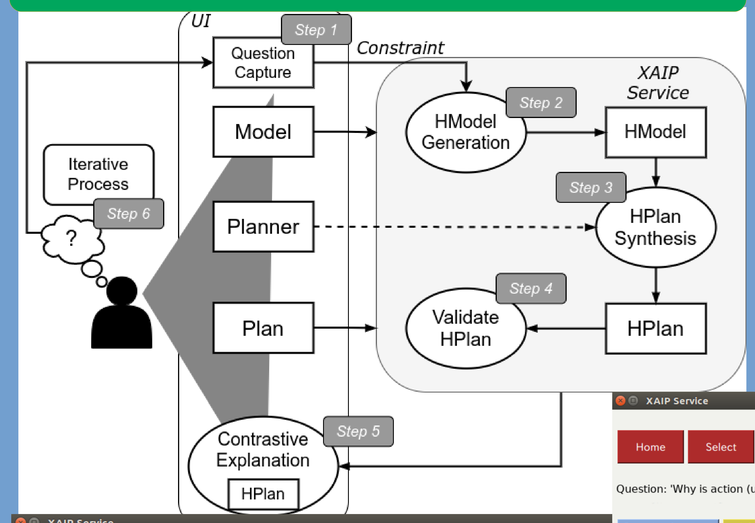
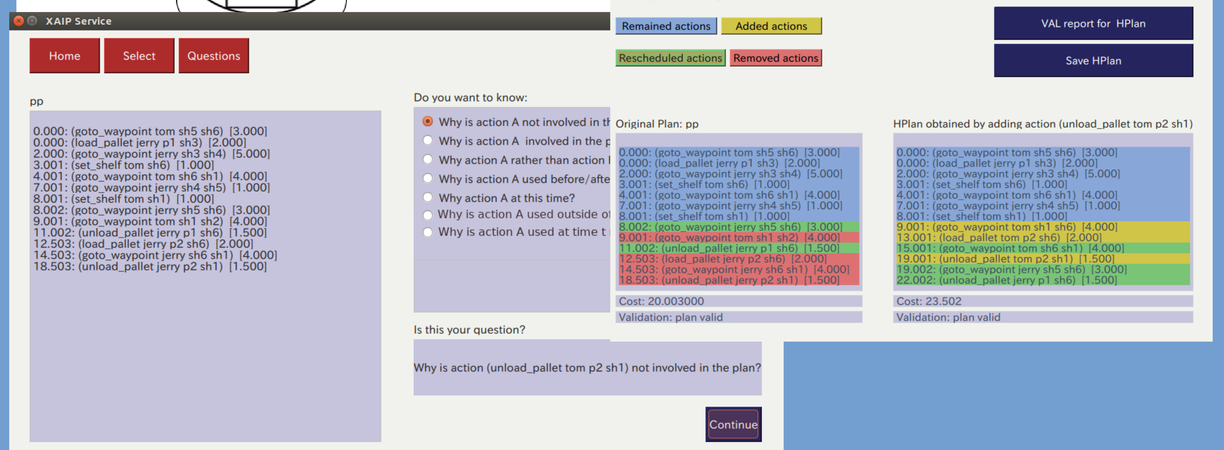
Contrastive Explanations of Plans Through Model Restrictions
Journal of Artificial Intelligence Research, Volume 72, 2021, 533-621

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

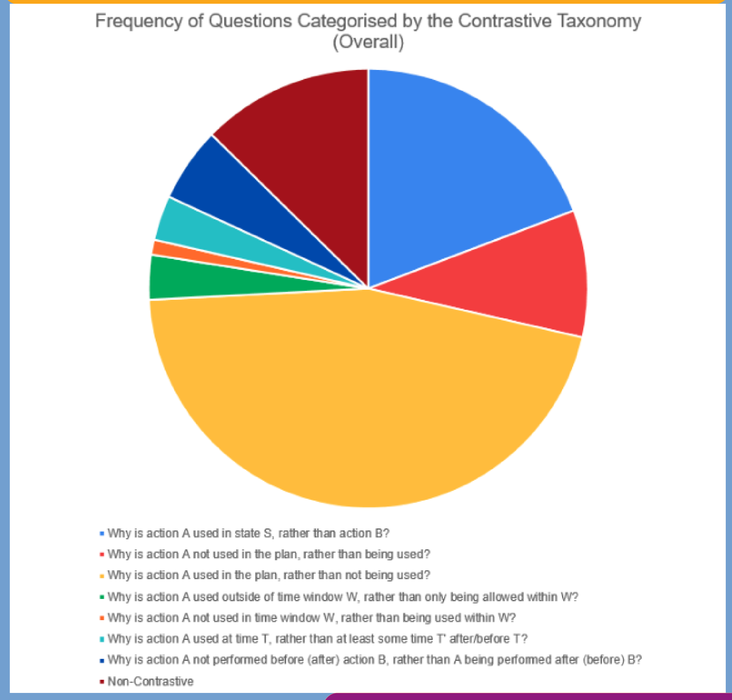
In automated planning, the need for explanations arises when there is a mismatch between a proposed plan and the user's expectation. We frame Explainable AI Planning as an iterative plan exploration process, in which the user asks a succession of contrastive questions that lead to the generation and 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





Why Contrastive Explanations?



Example Compilation

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

where:

- $Ps' = Ps \cup \{can\_do\_a, not\_done\_a\}$
- $As' = \{o_a, o_{\neg a}\} \cup As \setminus \{o\}$
- $arity'(x) = arity(x), \forall x \in arity$
- $arity'(can\_do\_a) = arity'(not\_done\_a) = arity'(o_a) = arity'(o_{\neg a}) = arity(o)$
- $I' = I \cup \{ground(not\_done\_a, \chi)\}$
- $G' = G \cup \{ground(not\_done\_a, \chi)\}$
- $W' = W \cup \{\langle lb, ub, ground(can\_do\_a, \chi) \rangle\}$

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