How Humans Explain the Difference in the Quality of Plans – A User Study

Benjamin Krarup[1], **Amanda Coles**[1], **Dancheng Gao**[1], **Derek Long**[1], **David E. Smith**

[1]King's College London, UK.
benjamin.krarup@kcl.ac.uk, amanda.coles@kcl.ac.uk, derek.long@kcl.ac.uk, david.smith@psresearch.xyz

Abstract

Recent advances in plan explanation have used abstractions to produce explanations. We consider the task of explaining why there is a difference in the quality of plans produced for a planning problem, Π, and the same problem constrained in some way, $\Pi + c$. The method involves abstracting away details of the planning problems until the difference in the quality of plans they support is minimised. It is not known whether humans use abstractions to explain these differences, and if so, what types of properties these abstractions have. We present the results of a qualitative user study investigating this. We tasked participants with explaining the difference in the quality of plans and found that users do indeed use abstractions to explain differences. We extract a set of properties that these abstractions satisfy, which can be used in automatic abstraction for explanation generation.

1 Introduction

Recent advances in plan-explanation have exploited abstractions (Göbeldecker et al. 2010; Sreedharan et al. 2019; Krarup et al. 2024). Abstraction is a process of simplification, where details of a problem are removed (Giunchiglia and Walsh 1992). In this work we consider abstractions to explain differences in the quality of plans. In a planning problem there is an initial state, I, a goal, G, and a set of actions, A, that prescribe the conditions under which they may be applied and the effects they have. For example, a task may involve using a delivery truck to deliver a number of packages to specific locations. The solution to a planning problem is a timestamped set of actions that transform the state I to a state satisfying G. An abstraction of a planning problem is a mapping to a new problem that is less constrained. The abstract planning problem admits all of the solutions to the original problem and more. Abstractions of the delivery example include the truck no longer consuming fuel, or carrying packages of any size.

We investigate, via a user study, to what extent humans use abstractions to explain the differences between solutions to similar planning problems. Having confirmed that they do, we extract a number of properties from the abstractions that can be used to aid in the design of automatic abstraction for human-like-explanation generation. We make a number of design recommendations based on our findings.

2 Related Work

The adoption of AI planning systems requires the ability to explain their behaviour. Fox et al. 2017 highlight the importance of contrastive 'why' questions in plan explanation, describing variants of these questions and possible responses. Chakraborti et al. 2017 approach explanation as model reconciliation, that is, explanation depends on demonstrating differences between the agent's and the human's models of the planning problem. Krarup et al. 2021 produce contrastive explanations by constraining planning problems to adhere to user specified contrast cases. There has also been interest in providing explanations in path/motion planning (Almagor and Lahijanian 2020; Pozanco et al. 2022).

Abstraction has an established role in problem solving, for example using heuristics based on abstracted (relaxed) problems to guide search. However, the use of abstraction in generating plan explanations is relatively recent. Gobeldecker et al. 2010 focus on finding changes to the initial state that would make a planning problem solvable, and provide an algorithm to produce these 'excuses' in reasonable time. Sreedharan et al. 2019 use abstractions of predicates to simplify planning problems to help explain unsolvability. Eifler et al. 2022 explain why some set of soft goals cannot be achieved through constraint relaxations. Krarup et al. 2024 utilise abstractions to explain the difference in quality of plans. They define an abstraction of a planning problem and search a space of abstractions until one is found that makes the plans equi-cost. This abstraction is then used as the basis for explanation. Sreedharan et al. (Sreedharan, Srivastava, and Kambhampati 2021) and Vasileiou and Yeoh 2023 use abstraction for generating personalised explanations whose level is based on a human's expertise with the task. Brandao et al. 2021 explain why some path is optimal rather than another by abstracting the navigation graph of the path.

Some of these abstraction based approaches are evaluated via user studies testing whether the explanations are satisfactory; but they are not evaluated in direct comparison with one another. It also remains to be demonstrated that using abstraction for explanation corresponds to the way humans explain plans. One purpose of this study is to determine whether humans generally use abstraction in explain-

ing plan quality differences, to motivate and support its use in future explanation-generation work. Existing approaches do not take into account the properties of abstractions that make them useful in explanation to humans. The second purpose of this study is to discover such properties, to help in generating explanations automatically.

Literature from the social sciences has supported the claim that humans use abstraction for explanation in a variety of contexts (Giunchiglia and Walsh 1992; Hitchcock and Woodward 2003; Miller 2019). However, to our knowledge, this is the first exploratory study to determine what explanations humans produce to explain plan quality differences.

3 Explanations Via Abstraction in Planning

A planning problem is a tuple, $\Pi = \langle I, G, A, M \rangle$, where I is the initial state, G is the goal to be achieved, A is the set of possible actions and M is the metric function that can be used to evaluate the cost or *quality* of a plan. Actions have preconditions restricting the states in which they may be executed, and effects describing the changes they make to states. Actions also have constraints on their duration: the maximum/minimum time they take to execute. The metric function may be based on plan duration (makespan), or a sum of costs, which might involve things such as energy consumed, heat generated, or risk. The solution to a planning problem, Π, is a plan, $\pi = \langle a_1, \ldots, a_n \rangle$, which is a collection of actions, $a_i \in A$, each with a specified start time relative to the beginning of the plan and a specified duration. Executing the plan will transform the initial state, I, to a goal state g, such that $G \in g$. The cost of the plan evaluated using the metric function is $M(\pi)$. We assume that the cost of plans is to be minimised. We also make use of a function, $D(\pi_1, \pi_2) = |M(\pi_1) - M(\pi_2)|$, which returns the difference in quality of any two plans, π_1 and π_2. If, for any two problems, Π_1 and Π_2, Π_2 is unsolvable, then $D(\pi_1, \pi_2) = M(\pi_1) + \gamma$ where γ is a suitably large overestimate for the worst case quality of a solution for Π_2. If $D(\pi_1, \pi_2) = 0$, we call π_1 and π_2 *equi-cost*.

States can be represented as subsets of a finite universe of propositional fluents, P, and a valuation of numeric variables V. The initial state is a subset of propositional fluents, $I \subseteq P$, that is initially true and an initial valuation of V. The goal is represented as $G \subseteq P$, and numeric conditions over variables in V. Action preconditions are sets of fluents and numeric conditions that must be true for the action to be performed. Effects are updates to the set of fluents and variables. Additionally timed-initial-literals (TILs) can make propositions true or false at specified times. Planning problems are formalized in this representation. However, participants in our study were given a text representation of the test problems, as we did not want to restrict the study to those familiar with planning modelling.

We consider explanations in the setting described by Krarup et al. 2024, as follows: given a planning problem, Π; a plan, π, for Π; a constraint, c, which π does not satisfy, and a solution plan, π^c, for $\Pi^c = \Pi + c$ (Π restricted to admit only solutions that obey c). We assume that there is a difference in the quality of π and π^c. A special case is where the problem Π^c is unsolvable and π^c does not exist. We seek to explain why there is a difference in the quality of π and π^c. We assume that an explanation of the form "the difference is because of the constraint c" is not helpful.

Planning problems are often constrained in this way in mixed-initiative settings: a planner is used to produce plans while a human adds constraints and preferences to the planning problem until they are satisfied with the result. Another example is in contrastive question answering (Krarup et al. 2021). Users can ask questions of the form, "Why A rather than B?", where A is a feature of the plan and B is some contrast case. To answer these questions the problem can be constrained so that the solution contains B rather than A. Explanations focus on the differences between these solutions.

An explanation of a discrepancy in plan quality should consist of elements of the problem (apart from c) that cause the discrepancy. A planning problem, Π_α, is an abstraction of Π if every solution of Π is a solution of Π_α. If the plans π and π^c are not of the same quality, but under the abstraction α, the plans π_α and π^c_α are the same quality, then we can say that α is a *cause* of the difference in quality of the plans for Π and Π^c. These causes can be found by abstracting away α from both Π and Π^c, until $D(\pi_\alpha, \pi^c_\alpha) = 0$.

As an example, consider a delivery problem Π and the contrastive question *"Why did you use Truck 1 instead of Truck 2?"*. This generates a constraint c that Truck 2 must be used in the plan. The problem Π^c is solved resulting in a longer plan using Truck 2. A *descriptive* explanation might be 'Truck 2 has to take a longer route'. A causal explanation can be found by searching over abstractions (removing action conditions, durations etc.) and discovering that abstracting a weight limit condition on crossing a bridge admits equi-cost plans using Truck 2 or Truck 1. This abstraction is a *cause* of the difference between the plans.

A user may question decisions made in a plan, π, for Π and explanations may uncover un-modelled preferences or goals, but we assume that the desire when performing abstraction is to do minimal damage to the the original planning problem, Π. With this assumption in mind, we introduce two quality measures for abstractions, degree of convergence and degree of perturbation. The degree of convergence measures the degree to which an abstraction causes the plans produced for the abstracted problems to converge. The degree of perturbation measures the degree to which an abstraction causes the plan π_α produced using the abstracted original problem to change with respect to the plan π for the original problem. We use plan quality as a proxy for these measurements. More specifically, the degree of convergence is defined by:

$$Conv(\Pi, \Pi^c, \alpha) = 1 - \frac{D(\pi_\alpha, \pi^c_\alpha)}{D(\pi, \pi^c)}$$

The degree of perturbation is defined by:

$$Pert(\Pi, \alpha) = \frac{D(\pi, \pi_\alpha)}{M(\pi)}$$

We assume that Π is always solvable, so we always have a plan π: given that our intention is to explain the difference in quality between a plan and a reference plan, π, if we do not have π we do not have a question to answer. Finally we

note the denominator is larger than the numerator for $Pert$ and $Conv$. For $Conv$ this holds because if π^c does not exist, i.e. Π^c is unsolvable, then $D(\pi, \pi^c) = M(\pi) + \gamma$, and we assume γ is sufficiently large so that $M(\pi^c_\alpha) < \gamma$. This holds for $Pert$ and for $Conv$, where both plans exist, due to the nature of abstraction. Any plan π for Π is a plan for Π_α, so there exists a plan π_α for Π_α that is at least as good as π, i.e, $M(\pi_\alpha) \leq M(\pi)$. Also note that $Pert$ does not take the constrained problem as an argument; the degree of perturbation for an abstraction is strictly a measure of how much the abstraction damages the original problem, Π.

We are interested in four cases for Π, Π^c, and α:

1. $Conv(\Pi, \Pi^c, \alpha) = 1$ and $Pert(\Pi, \alpha) = 0$. These abstractions seem most suitable for explantation since they fully explain the difference in the quality of π and π^c and they do not perturb the original problem, Π. We call these abstractions **fully convergent** with respect to Π and Π^c and **non-perturbing** with respect to Π.
2. $Conv(\Pi, \Pi^c, \alpha) = 1$ and $Pert(\Pi, \alpha) > 0$. These abstractions still fully explain the difference in the quality of π and π^c but perturb the original problem, Π. These abstractions are **fully convergent** and **perturbing**.
3. $0 < Conv(\Pi, \Pi^c, \alpha) < 1$ and $Pert(\Pi, \alpha) \geq 0$. These abstractions partially explain the difference in the quality of plans and may perturb the original problem Π. These abstractions are **partially convergent** and may or may not be **perturbing**.
4. $Conv(\Pi, \Pi^c, \alpha) = 0$ and $Pert(\Pi, \alpha) \geq 0$. These abstractions do not reduce the difference in quality of plans at all and therefore cannot explain the difference, we call these **non-convergent** or **irrelevant**.

There are further cases e.g. **fully perturbing** abstractions, which abstract problem so that an empty plan is valid (e.g. by removing all goals), but this is not useful for explanation.

Current work utilises blind search to find these abstractions. There is no consideration of which possible abstractions are useful in producing more satisfactory explanations. One aim of this study was to determine what types of abstractions are used by humans in explanation. There may be many fully convergent abstractions so it is useful to know which are more natural to humans in order to inform abstraction selection in automatic generation of explanations.

4 Study Design

We designed a user study [1] to investigate how people explain plan quality differences. We considered several hypotheses, namely that the participants' explanations correspond to:

- **(H1)** Abstractions of the problem.
- **(H2)** Abstractions that are fully convergent.
- **(H3)** Abstractions that, when they are fully convergent, they are also non-perturbing.
- **(H4)** A single abstraction.

And:

- **(H5)** Explanations are formed from information obtained from the planning problem, the question posed, the original /constrained plans and through a process of abstraction and additional reasoning.
- **(H6)** Participants produce causal explanations rather than descriptive explanations.

We hypothesise that humans use abstractions to explain the difference in the quality of plans (H1). We hypothesise that these abstractions will fully explain the difference in the quality of plans, they will be fully convergent (H2). Furthermore, we hypothesise that, if the explanations contain causes (are minimally perturbing), humans will explain with respect to the ground truth baseline planning problem, and therefore the abstractions will be minimally perturbing (H3).

We hypothesise that humans give simple explanations: namely they will give only the causes necessary to explain the difference. We expect this to lead to explanations comprising single abstractions that explain the difference (H4).

We hypothesise that explanations humans produce to explain differences in plan quality will contain information that is readily available from the description of the planning problem, the question asked, the given plans, and other information inferred from these by the user, via causal or contrastive reasoning and a process of abstraction (H5). This will help to shine a light on the possibility of automatically generating explanations given this information, which is often available in model based planning scenarios.

Finally, we hypothesise that humans give causal (Lewis 1974) rather than descriptive explanations (H6). Causal explanations give reasons *why* there is a difference in plan quality; whereas descriptive ones simply describe what the differences are. Other work distinguishes between these types of explanations as answers to "why?" and "what?" questions (Miller 2019).

The hypotheses represent expectations and are used to discuss the results of the study. We also provide the outcomes of statistical tests, along with observations about the qualitative data, to show support for these hypothesis. However, the objective of this study was not to confirm or deny these hypotheses. This confirmation testing is left to future work.

4.1 Methodology

In order to explore our hypotheses, we designed a qualitative study in the form of a questionnaire. We recruited 20 participants using Prolific. We selected a sample size of 20 as this has been shown to cause data saturation in qualitative studies (Nielsen 2000; Faulkner 2003). The participants were recruited from the UK, Ireland, the USA, Australia, and Canada, whose primary language is English. The participants were aged between 18 and 57, 11 identified as female while 9 identified as male. Occupations declared by the participants included engineer, software engineer, data scientist, carer, therapist, and unemployed. Each participant was presented with four different planning problems in a random order. They were given a description of the planning problem in natural language. We ensured that there was no extra information in the description that would not be present in the planning problem. However, we described the problem

[1] The study, planning problems, results, and analysis can be found at https://github.com/BenKrarup/PlanAbstractionStudy

as if the participant were the modeller, so they had all of the environmental information necessary to model the problem. We also provided the participants with a visual description of the task. For each problem we presented users with the optimal plan, in natural language and as a visual diagram, to solve the problem. The participants were then tasked with answering two questions for each problem. For each question the users were asked to imagine that they had applied a specific constraint to the problem such that the optimal solution was no longer valid, and the new solution was of worse quality or the problem was no longer solvable. If solvable, they were presented with the worse quality plan. The participants were then asked to give an explanation for why, given the constraint that was added to the problem, the problem was unsolvable or the new plan was of lesser quality.

Delivery Task A driver must deliver a package of meat to a butcher and cereals to a grocer. It takes 10 minutes to drive between the depot and grocer; 15 to drive between the grocer and butcher; and 20 to drive across the bridge between the depot and butcher. It takes 1 minute to load packages into a truck, and 1 for the driver to board the truck, and 2 to unload and deliver packages. If the meat package is unrefrigerated for more than 21 minutes it will spoil. There is a weight limit on the bridge which means that only small trucks can cross.

In this task, we asked: "Explain why it takes longer to solve the problem using the unrefrigerated small truck T1 rather than the refrigerated small truck T2?" and "Explain why the problem becomes unsolvable when only unrefrigerated large truck T3 is used instead of the unrefrigerated small truck T1 or refrigerated small truck T2?"

Satellite Task A satellite must take infrared or visible images of planets, stars, and phenomena. The satellite has two imaging instruments, I1 and I2, that support different imaging modes. The initial state of the task is shown in Figure 1. Instruments must be turned on and calibrated before use, and only one instrument can be turned on at a time. The goal is to take infrared images of Star 2, Planet 1, Planet 2, and Phenomenon 2; and a visible image of Phenomenon 1.

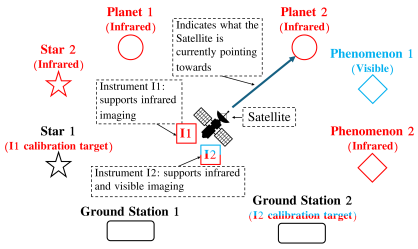


Figure 1: Initial state of the Satellite Task.

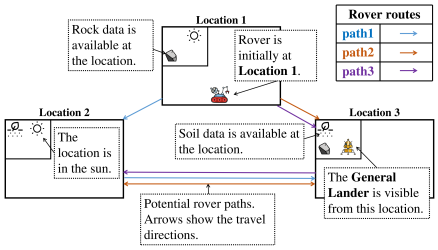
In this task, we asked: "Explain why it takes longer to solve the problem using instrument I1 to take the infrared image of Phenomenon 2 rather than instrument I2?" and "Explain why the problem becomes unsolvable when only instrument I1 can be used instead of instrument I2?"

Building Task Two locations are connected by a road: Resource Land where you can gather wood and iron; and Empty Land. Gathered resources can be used to build vehicles and, on Empty Land, small or big houses. Building a big house requires 1 more iron than a small house. A vehicle (either a cart or a train) must be built to transfer resources between Resource Land and Empty Land. A train has greater capacity and travels faster than a cart, but requires more time and resources to build. A rail must also be built to use a train. A visual description of the task (similar to the other tasks) was presented to the study participants. The goal is to build 2 houses of any type in the shortest possible time.

Here we asked: "Explain why it takes longer to solve the problem using the train rather than the cart?" and "Explain why it takes longer to solve the problem when building 2 big houses rather than 2 small houses?"

Rover Task A rover must collect rock data from Location 1, soil data from Locations 2 and 3, and communicate this data back to the Lander. There are three paths between these locations and the rover can only communicate data from a location where the Lander is visible. The initial energy level of the rover is 0 and each action consumes a certain amount of energy. The rover can only recharge at locations that are in the sun. The initial state is presented in Figure 2. The goal is to complete all 3 tasks in the shortest possible time.

Figure 2: Initial state of the Rovers Task.



Here, we asked: "Explain why it takes longer to complete the tasks when path2 is used instead of path1?" and "Explain why the problem becomes unsolvable when we use path3?"

5 Qualitative Methodology to Analyse Data

To explore our hypotheses we collected qualitative data. In this section we outline the methods and procedures we used to analyse the qualitative data. We followed a procedure of open and then focused coding (Corbin and Strauss 1990). This mainly consisted of data coding through labelling of key features and themes of the explanations users produced.

We analysed the explanations users produced in three different ways. We first extracted the abstractions used to produce the explanation. We then labelled the data based on the likely source of each part of the explanation. Finally, we categorised the explanation as causal or descriptive.

5.1 Extraction of Abstractions

In order to test hypotheses H1 - H4 we extracted the abstractions that correspond to participants' explanations. We did this through the use of the constructed PDDL2.2 (Edelkamp and Hoffmann 2004) planning problems, and the process of abstraction for extracting causes, described in Section 3.

For each explanation we identified the reason cited by the participants for the difference in the quality of the solutions that were presented to the participants. These reasons were properties contained within the task descriptions presented to the participants. We abstracted away the accompanying property in the PDDL problem to determine if it was a cause, and calculated the degree of convergence and perturbation.

These abstractions were often preconditions of actions. For example, from the explanation for question two of the Rovers Task, "Because the rover can only communicate the data when it is at location 3, where the lander is visible. Path 3 only gets to collect the data at location 2 but can never transmit it without returning to location 3.", two precondition abstractions were extracted. The first is the precondition that ensures that the rover can only communicate data when the lander is visible. This was extracted due to the participant citing the precondition in their explanation, "the rover can only communicate the data when it is at location 3, where the lander is visible", and this is a cause. The second is the precondition that allows the rover to navigate between certain locations. Again, the participant cited this precondition, "can never transmit it without returning to location 3." Here the participant notes that because we cannot return to location 3 we cannot transmit the data, which is also a cause.

We also found abstractions that were action durations. For example, from the explanation for question one of the Building Task, "although the train is faster, it takes time to build the train and rails", the abstraction of two actions' durations was extracted. The first is the duration of the action to build the train. This was extracted because the participant cited the time taken to perform this action as a reason, "it takes time to build the train". Similarly, the second abstraction was the duration of the action to build the rails. These are both causes.

Although we have been referring to abstractions that reduce the difference in the quality of plans as 'causes', we still extracted abstractions from descriptive explanations. For example, the explanation for question one of the Rovers Task "When path 2 is used, the rover must navigate a total of 3 times (15 minutes total), while in path 1 the rover only navigates twice (2 times). So this extra navigation of going back and forth to location 2 adds the 5 minutes." is descriptive. The participant does not give a causal reason for why the rover must navigate more times. They describe that this is the case in the plans, and that this takes longer. However, an abstraction can still be extracted. The participant cites the extra navigation as the reason, abstracting away the duration of the navigate action does reduce the difference in the quality of the plans. This is not a causal explanation because the participant does not give the reason for the extra navigation. Instead, they give the reason the quality was different: presence of an extra navigation step in the plan.

Two independent coders extracted abstractions from each response. Inter-rater reliability was assessed using Cohen's kappa for each abstraction type, with separate κ coefficients computed for the presence/absence of each abstraction across relevant responses. This indicated consistent classification between coders. Coding conflicts were adjudicated by consensus.

5.2 Source of the Explanations

We used a method of qualitative data coding through labelling to determine the source of each phrase or part of the participants' explanations, to test hypothesis H5. By source, we mean where the knowledge needed to produce the explanation was available. We identified six different sources of knowledge used to produce these explanations and categorised the explanations by these sources through labelling. The identified sources were: the problem description, the question posed, the original and constrained plans, and information obtained through a process of abstraction, by contrastive, or by causal reasoning. The final source category was any extra information where the source was unclear.

We labelled a phrase in the explanation as from the original or constrained plans if it referred to actions that were present in those plans. The source of a phrase was categorised as from the original or constrained problems if the phrase referred to information that was available in these descriptions. We labelled a phrase as from the question information if the information was available in the question that the participants were tasked with answering. We labelled a phrase as from a process of abstraction if they referenced any information available in the abstracted problem and plans that we generated through the abstractions extracted from the explanation. The participants did not have this information explicitly, but similarly as described in Section5.1 through the process of abstraction described by Krarup et al. 2024 we generated the abstracted problems and plans to see if it was a source. A phrase was labelled as from contrastive reasoning if the participant clearly formed some conclusions based on contrasting information presented in the plans, or through some hypothetical scenario that was not presented to them, for example, reasoning that a certain path to the goal would be preferable to another. We labelled a phrase as from causal reasoning if the participant clearly had to reason about some causal information to form some conclusion, for example, reasoning that some condition must be satisfied in order to perform some action in the plan.

We illustrate our method of data coding on an explanation for question one of the Delivery Task: "It takes longer to use the unrefrigerated truck because the meat will spoil after 21 minutes. This means that the truck must first go to the butcher from the depot and this is a longer journey (20 minutes) than the refrigerated truck's route which involves going straight to the grocer." The information in the phrase, "It takes longer to use the unrefrigerated truck because the meat will spoil after 21 minutes", was available in the original and constrained plan, the domain information, and through some contrastive reasoning. The original plan used the refrigerated truck, the constrained plan used the unrefrigerated truck and took longer. The meat spoiling after 21 minutes is in the description of the task, and through some contrastive reasoning the participant can deduce from these three sources of information that "it takes longer to use the unrefrigerated truck" because if the unrefrigerated truck used the same route as in the original plan then "the meat will spoil after 21 minutes". The source of the phrase, "This means that the truck must first go to the butcher from the depot and this is a longer journey (20 minutes) than the refrigerated trucks route which in-

volves going straight to the grocer" was the original plan, the constrained plan, and contrastive reasoning. This explanation contrasts the original plan with the constrained plan and notes that the latter is longer.

5.3 Causal vs. Descriptive Explanations

We categorise an explanation as causal if it includes some causal information explaining *why* there is a difference in the quality of two solutions and as descriptive if no causal information appears in it but, instead, it focuses on *what* are the differences between the solutions.

For example, the explanation for question one of the Satellite Task, "because you need to take both types of image and I1 can only take 1 type so both have to be used, so both have to be calibrated, adding extra time.", was categorised causal because it gives a reason that the constrained plan takes longer. It correctly asserts that I1 can only take one type of image, while the goal requires two different types of images. Therefore I2 must be used taking longer as this must be turned on and calibrated. In contrast, the explanation, "Because more steps are involved as two instruments are being used", was categorised as descriptive. It describes the difference between the two plans: in the constrained plan, two instruments are used instead of one. It does not give a cause for two instruments being used, and why this makes the solution take longer.

Two independent coders classified each response as either causal or descriptive. Inter-rater reliability for this binary coding was assessed using Cohen's kappa, $\kappa = 0.63$, and percentage agreement, 0.89%, indicating substantial agreement between coders. Conflicts in coding were adjudicated by consensus.

6 Results and Analysis

From the 8 questions for the 4 problems we presented to the 20 participants, we received a total of 160 explanations. The majority of these explanations were a couple of sentences long. The longest explanation given was 88 words while the shortest was 7 words. None of the explanations produced by participants had to be dismissed due to illegibility. We did not observe a notable difference in the explanations given by participants based on their backgrounds. Of these 160 explanations, 123 were causal explanations and 37 were descriptive explanations (H6). Performing a binomial test with $p = 0.5$ as the null probability of an explanation being causal the results are significant at $p = 0.01$, supporting H6. In the rest of this section, we will present the results of our analysis of the participants' explanations for the purpose of evaluating our hypotheses presented in Section 4. For all binomial tests we use a null probability $p = 0.05$ as a non-informative baseline. Although it may seem some outcomes are structurally more likely, e.g. explanations contain abstractions, no prior empirical evidence exists to inform this null probability and we are testing this for the first time. Therefore we did not want to allow any imagined prior probability to influence the results. Each of our classifications are binary and balanced by design. Therefore, we assume each outcome is equally likely.

6.1 Abstractions

From the 160 explanations participants produced, a total of 265 abstractions were extracted including 53 different abstractions. We reached data saturation: no new kinds of explanations were being produced from the study and no new abstractions were being extracted. We extracted abstractions from 151 of the explanations produced. We could not extract an abstractions for 9 of the explanations based on our coding scheme. Of the explanations produced, 94% corresponded to abstractions of the problem presented (H1). Performing a binomial test with $p = 0.5$ as the null probability that explanations correspond to abstractions of the problem was significant at $p = 0.01$, supporting H1.

The abstractions extracted are shown in Table 1. The table shows the type of abstraction that was extracted, the abstraction, the code that was assigned through coding, the effect of the abstraction, and the number of explanations that mentioned this abstraction. The code was extracted through thematic analysis of the explanations. Then, through the use of the PDDL problem and abstraction description in Section 3 we assigned the code with its corresponding abstraction as well as the type of the abstraction. Six types of abstractions were present in the explanations. Codes that mentioned: conditions or properties of objects were classified as precondition abstractions; the time for actions to execute as duration abstractions; time constraints as timed-initial-literal (TIL) abstractions; numeric conditions as function abstractions; ordering on the execution of actions as order abstractions; and the constraint that was imposed by the question given in the study as a constraint abstraction.

Participants predominantly referenced precondition (114) and duration abstractions (101). We conjecture that this is because action conditions and durations are a feature of most planning problems, including those in the study. Each of these problems had preconditions that could not be achieved or action durations that were causes of differences in the quality of the solutions presented to the participants. Only one type of abstraction, precondition, was extracted from the explanations for question 2 of the Satellite Task, probably because of the nature of the task description. The constraint added to the Satellite Task causes it to become unsolvable due to the instrument I1 not being able to take visible images. Abstracting this precondition makes the problem solvable. Participant explanations mainly centred on this fact. Function (27) and TIL abstractions (21) were the third and fourth most common. TIL abstractions were only extracted from explanations for the Delivery Task, probably because a time constraint is a crucial feature of the task (the meat spoiling after 21 minutes). Function abstractions were present in each of the problems. Finally, there was one occurrence of each of the constraint and order abstractions. The constraint abstraction is not one we believe to be useful, as explanations containing them answer the question posed with the negation of the question. This data supports this claim.

Table 1 shows the effect the abstractions have on the quality of the plans. CN, indicates solutions to abstractions of both the original and constrained problems are equi-cost, and these are also equi-cost with the original plan, so they are fully convergent and non-perturbing. C, indicates that

the abstractions were fully convergent but perturbing. P, indicates abstractions are partially convergent and may or may not be perturbing. I indicates the abstraction is irrelevant.

The majority of abstractions, 127, were CN abstractions. 32 abstractions were C abstractions. 58 of the abstractions were partial and 48 were irrelevant. Question 2 of the Building Task had the most explanations with partial abstractions. This may be because participants correctly identified that more time would be needed to gather the resources for big houses, but did not account for the extra time to load and unload them. Question 1 of the Satellite Task had the most explanations with irrelevant abstractions. This is likely due to participants believing that the quality difference was caused by turning on, calibrating, and turning off the extra instrument, but, these actions can be done in parallel with other actions so was not the reason for the difference in quality.

Of the 265 abstractions extracted 60% were fully convergent (H2). Performing a binomial test with $p = 0.5$ as the null probability that an extracted abstraction removes the difference in the quality of the solutions, these results are significant at $p = 0.01$, supporting H2.

Of the 159 abstractions extracted that were fully convergent 79.9% were also non-perturbing (H3). Performing a binomial test with $p = 0.5$ as the null probability that when an extracted abstraction was fully convergent they were also non-perturbing, these results are significant at $p = 0.01$, supporting H3.

Multiple abstractions were extracted from each explanation and each was evaluated based on its individual effects. We hypothesised that each explanation would correspond to only one abstraction (H4). Of the 151 explanations containing an abstraction, 71 explanations corresponded to one abstraction, 43 to two, 30 to three, 6 to four, and 1 to five. We tested H4 using a binomial test with p = 0.5 as the null probability of an explanation containing only one abstraction. The probability of observing 71 single-abstraction explanations out of 151 was p = 0.26, which is not statistically significant.

Finally, we address the question of whether participants were indeed understanding the problem and utilising abstraction to produce their explanation or were simply doing feature matching on plans. There is considerable empirical evidence that contradicts the use of a simple feature matching approach on plans. Firstly, participants most frequently produced causal explanations; yet the information necessary to produce causal explanations is not present in the plans. Secondly, three of the constraints applied to the planning problems presented caused them to become unsolvable, therefore there was no constrained plan for the participants to compare to. We observed no noticeable difference in the explanations produced by the participants for these problems. Finally, as we will see in the next subsection, 80.7% of the information in explanations, we recognised through coding, came from sources other than the plans. This demonstrates participants made considerable use of information from other sources to generate explanations.

6.2 Sources of Explanations

The sources of information for participant's explanations are shown in Figure 3. These correspond to the nine sources identified in Section 5. This distribution was created using the number of characters in the component of the explanation from each source. Of the sources classed as contrastive and causal reasoning, explanations contained more information available through contrastive reasoning (28.8%) than causal reasoning (12.3%). From the other sources of information, explanations contained the most information from the problem description (18.6%). Explanations used little information from the question (0.3%). It was not possible to assign a source for 11.7% of the information contained in the explanations. This was due to difficulty in determining where the information was available. 88.3% of information in the explanations was available from the problem description, the original and constrained plans, the question posed, through a process of abstraction, and contrastive and causal reasoning (H5).

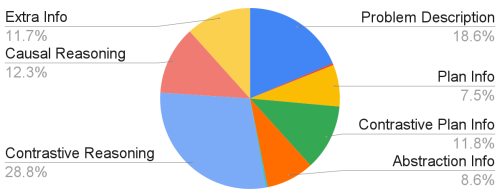


Figure 3: Distribution of the sources of the explanations produced by participants.

7 Conclusions and Recommendations

Humans heavily use abstractions to explain the reason for the difference in the quality of plans (H1). Many abstractions completely, rather than partially, explain the differences (H2), and when they did, caused the constrained problem to produce solutions equi-cost to the original solution (H3). A substantial fraction of explanations that humans produce correspond to only a single abstraction (H4). The explanations produced by participants were largely constructed from information available in the task description and solutions they support (H5). This indicates that human-like-explanations may be generated from information in these sources. A substantial proportion of explanations that participants produced were causal rather than descriptive (H6).

Although abstraction is a popular approach to explanation, until now there has been limited evidence that humans utilise abstraction to explain plan quality differences. Support for H1 and H6 provide foundational support for the use of causal explanation via abstraction in this setting.

Prior work has not considered what abstractions produce the best explanations. From hypotheses H2 to H5, We can recommend that for human-like-explanations, abstractions should completely explain differences, cause the constrained problem to produce solutions equi-cost to the original solution, be minimal, and the explanation itself should contain information from the task description and solutions. In Krarup et al 2026, we use this information to help guide search for finding abstractions that satisfy these properties, thus allowing us to automatically generate better explanations of plan quality differences.

Type	Abstraction	Code	Effect	Count
		Delivery Task Question 1		
Pre	Refrigeration	T2 is refrigerated/T1 is unrefrigerated	CN	12
	At	Takes two paths instead of one	C	1
Dur	Drive	Longer travel time	CN	16
TIL	Spoiled Meat	Meat will spoil after 21 mins	CN	11
Fun	Number of paths	It takes two paths instead of one	I	1
		Delivery Task Question 2		
Pre	Refrigeration	T3 is unrefrigerated	CN	10
	CanTraverse	T3 is too heavy	P	16
Dur	Drive	Meat cannot be delivered on time	CN	5
TIL	SpoiledMeat	Meat will spoil after 21 mins	CN	10
Fun	Distance	T3 has to take the long road	CN	1
		Satellite Task Question 1		
Precondition	Can take image	I1 does not support visible mode	CN	7
	Power available	Only one imaging instrument can be turned on at a time	I	4
	Calibrated	Both instruments have to calibrated	C	4
	Turned on	I2 needs to be switched on after I1	I	1
	Calibration target	I1 and I2 have a different calibration target	P	1
Duration	Turn on instrument	Additional time needed to turn on instrument	I	9
	Calibrate target	Additional time needed to calibrate instrument	I	7
	Turn off instrument	Additional time needed to turn off instrument	I	5
	Turn to object	It takes time to turn the satellite to the target	CN	2
Constraint	Use I1	I2 should take all images, instead of I1	CN	1
Order	Order of positioning	The order of how the satellite points at objects	I	1
		Satellite Task Question 2		
Precondition	Can take image	I1 does not support visible mode	CN	18
		Building Task Question 1		
Precondition	Connected By Rail	A rail is needed to use the train	C	5
	Is Train	Must build the train, which takes more resources/time	C	5
Duration	Find Resource	More time spent gathering resources	CN	4
	Build Rail	More time spent building the rail	P	9
	Build Train	More time spent building the train	P	5
Function	Available Resources From Build Train/Build Rail	Requires more resources to build the train and the rail	CN	2
	Available Iron	It requires more iron to build the train and rail	CN	1
	Space In Train	The train's capacity is too small	P	2
		Building Task Question 2		
Precondition	Connected By Rail	A rail is needed to use the train	P	1
	Is Train	Although the train is quicker you have to build the train	P	1
Duration	Find Resource	More time spent gathering resources	P	3
	Find Iron	More time spent gathering iron	P	8
	Load	More time to load the resources	P	3
	Unload	More time to unload the resources	P	3
	Move Cart and Move Train	Transportation of resources takes more time	P	6
Function	Available Resources	More resources required	CN	2
	Available Resources From Build Big House	More resources required to build big houses	CN	6
	Available Iron	More iron required	CN	1
	Available Iron From Build Big House	More iron required to build big houses	CN	7
		Rovers Task Question 1		
Precondition	Lander Visible	Can only communicate to the lander from location 3	CN	4
	In Sun	There is no sun in location 3/can not reacharge	I	2
Duration	Navigate	Takes more time to navigate	CN	10
	Recharge	Takes more time to recharge	I	6
Function	Energy	Energy level	I	1
		Rovers Task Question 2		
Precondition	Lander Visible	Can only communicate to the lander from location 3	CN	13
	In Sun	There is no sun in location 3/can not recharge	I	8
	Can Traverse	Can not go back to location 3	CN	1
Function	Energy	Energy level	I	3

Table 1: Abstractions extracted from participants' explanations.

References

Almagor, S.; and Lahijanian, M. 2020. Explainable multi agent path finding. In *AAMAS*.

Brandao, M.; Coles, A.; and Magazzeni, D. 2021. Explaining path plan optimality: Fast explanation methods for navigation meshes using full and incremental inverse optimization. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 31, 56–64.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proc. International Joint Conf. on AI*.

Corbin, J. M.; and Strauss, A. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1): 3–21.

Edelkamp, S.; and Hoffmann, J. 2004. PDDL2. 2: The language for the classical part of the 4th international planning competition. Technical report, Technical Report 195, University of Freiburg.

Eifler, R.; Hoffmann, J.; and Frank, J. 2022. Explaining softgoal conflicts through constraint relaxations. In *31st International Joint Conference on Artificial Intelligence*.

Faulkner, L. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3): 379–383.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable Planning. *Proc. International Joint Conf. on AI-17 workshop on Explainable AI*, abs/1709.10256.

Giunchiglia, F.; and Walsh, T. 1992. A theory of abstraction. *Artificial intelligence*, 57(2-3): 323–389.

Göbeldecker, M.; Keller, T.; Eyerich, P.; Brenner, M.; and Nebel, B. 2010. Coming up with good excuses: What to do when no plan can be found. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Hitchcock, C.; and Woodward, J. 2003. Explanatory generalizations, part II: Plumbing explanatory depth. *Noûs*, 37(2): 181–199.

Krarup, B.; Coles, A.; Long, D.; and Smith, D. E. 2024. Explaining Plan Quality Differences. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, 324–332.

Krarup, B.; Coles, A.; Long, D.; and Smith, D. E. 2026. Finding Human-Aligned Abstractions Efficiently for Explaining Plan Quality Differences. In *Proc. International Conf. on Automated Planning and Scheduling*.

Krarup, B.; Krivic, S.; Magazzeni, D.; Long, D.; Cashmore, M.; and Smith, D. E. 2021. Contrastive Explanations of Plans through Model Restrictions. *JAIR*, 533–612.

Lewis, D. 1974. Causation. *The journal of philosophy*, 70(17): 556–567.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Nielsen, J. 2000. Why you only need to test with 5 users. In *Nielsen Norman Group, Nielsen*.

Pozanco, A.; Mosca, F.; Zehtabi, P.; Magazzeni, D.; and Kraus, S. 2022. Explaining preference-driven schedules: the expres framework. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 32, 710–718.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2021. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artificial Intelligence*, 301: 103570.

Sreedharan, S.; Srivastava, S.; Smith, D.; and Kambhampati, S. 2019. Why can't you do that hal? explaining unsolvability of planning tasks. In *International Joint Conference on Artificial Intelligence*.

Vasileiou, S. L.; and Yeoh, W. 2023. PLEASE: Generating Personalized Explanations in Human-Aware Planning. In *ECAI 2023*, 2411–2418. IOS Press.