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**[2]

[1]Department of Informatics, King's College London, London, UK
[2]PS Research, Los Altos Hills, CA, USA
benjamin.krarup@kcl.ac.uk, amanda.coles@kcl.ac.uk, dancheng.gao@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, leaving only what is relevant to the problem (Giunchiglia and Walsh 1992). In this work we consider abstractions to explain the 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) automatically generate answers to contrastive questions humans might pose about plans. 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 explain unsolvability of planning problems. 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 model and search a space of abstractions until one is found that makes two plans have similar cost. This abstraction is then used as the basis for explanation. Sreedharan et al (2021) and Vasileiou and Yeoh (2023) use abstraction for generating personalised explanations whose level is based on a human's knowledge or expertise of 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. However, it is not known if there is another technique that could be utilised to provide better explanations. These approaches 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. A purpose of this study is to determine whether humans generally use abstraction in explaining plan quality differences, to motivate and support its use in future explanation-generation work.

The abstraction approaches cited above typically use blind search. They do not take into account what properties abstractions have that may support or detract from their use in explanation. The second purpose of this study is to discover such properties, to help in the search for explanations.

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 in complex AI reasoning scenarios.

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 the plan. The actions have preconditions restricting the states in which they may be executed, and they have effects describing the states they lead to. Actions can also have a duration: the time required to execute. The metric function may be based on plan duration (makespan), or a sum of action 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 $g \in G$. The cost of the plan evaluated using the metric function is $M(\pi)$. We assume that the cost of plans is to be minimised.

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 preconditions over variables in V. Action preconditions are sets of fluents and numeric preconditions that must be true for the action to be performed. Effects are updates to the set of fluents and valuations to the state in which it is applied. Events are represented using timed-initial-literals (TILs) which make propositions true or false at specified times. Planning problems are formalized in this representation. However, participants in our study were presented with text-based 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', for $\Pi + c$ (Π restricted to admit only solutions that obey c). We assume that there is a difference in the quality of π and π'. A special case is where the model $\Pi + c$ is unsolvable and π' does not exist. We seek to explain why there is a difference in the quality of π and π'. 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 setting: a planner is used to produce plans while a human adds constraints and preferences to the model until they are satisfied with the resulting plan. 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', is an abstraction of Π if every solution of Π is a solution of Π'. If the plans π and π' are not of the same quality and the plans, π_α and π'_α, for the same problems abstracted with 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 $\Pi + c$ because the abstraction extends the plan space to include the equi-cost plans. Therefore, these causes can be found by abstracting away elements, α, of both of the planning problems, Π and $\Pi + c$, until $M(\pi) = M(\pi')$, we call α a *complete cause* for the differing cost plans. However, an abstraction can also be a *partial cause* if it reduces the difference in the quality of the solutions of the abstracted problems. If $|M(\pi) - M(\pi')| = n$ and α is applied to both Π and $\Pi + c$ and their solutions, π_α and π'_α, $|M(\pi_\alpha) - M(\pi'_\alpha)| = m$ and $m < n$, then α is a partial cause. An abstraction that causes the solutions of the problems to become equi-cost is called a *complete abstraction*, and one that reduces the difference in the costs is a *partial abstraction*.

As an example, consider a delivery problem Π and the contrastive question *"Why did you use Truck 1 instead of Truck 2 in the solution?"*. This generates a constraint c that Truck 2 must be used in the plan. The problem $\Pi + c$ 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, the bridge has a weight limit, meaning that Truck 2 must take a different route.

Current work utilises blind search to find these abstractions. There is no consideration of which of the 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 possible complete abstractions so it is useful to know which are more natural to humans to inform selection of abstractions for automatic generation of explanations.

4 Study Design

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

- **(H1)** Abstractions of the problem.

- **(H2)** Abstractions of the problem that remove the difference in the quality of the solutions to the original and the constrained problems.
- **(H3)** Abstractions that cause the constrained problem to produce plans similar in quality to the original plan.
- **(H4)** A single abstraction.

And:

- **(H5)** Explanations will be formed of information in the task description, with some additional reasoning.
- **(H6)** Participants produce causal explanations rather than descriptive explanations.

We hypothesise that humans use abstractions in order to explain the difference in the quality of plans. They give explanations in terms of causes that, if one were to imagine did not hold, then the difference in the quality of the plans would be minimised. This is the basis of abstraction as defined in Section 3. We hypothesise that these abstractions will explain the total difference in the quality of plans. So, humans will give complete explanations, rather than partial explanations. H1 and H2 capture these hypotheses.

We speculate humans explain the difference in the quality of plans with respect to the original plan presented. Given a planning problem, Π, a constraint, c, a plan for Π, π, and a plan for $\Pi + c$, π', where $M(\pi) < M(\pi')$, if there exists an abstraction, α, such that when the abstraction is applied to both Π and $\Pi + c$ their solution plans, π_α and π'_α, are equi-cost, and these plans are also equi-cost to the original plan π, then we believe this is a desirable property of abstractions for explanation. The original plan presented to the human is valid and optimal. Therefore, π'_α is also an optimal plan, and likely one that a human would accept as reasonable. H3 captures this hypothesis.

We hypothesise that humans give simple explanations. Hypothesis H4 is aimed at testing this. We postulate that when humans explain differences in plan quality they will give only the causes necessary to explain the difference. We believe this will manifest as human explanations consisting of only one abstraction that explains the difference.

We hypothesise that the explanations humans produce to explain differences in plan quality will contain information available in the description of the planning problem, with some additional reasoning. We say additional reasoning as an umbrella term for causal or contrastive inference that allows humans to make conclusions from information about the structure of problems. This hypothesis allows us to determine that we can likely automatically create explanations with a description of the planning problem, and through some additional reasoning. Hypothesis H5 tests this.

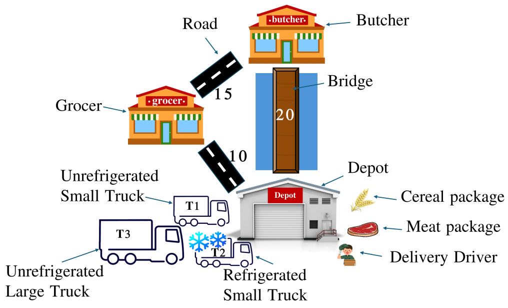
Finally, we hypothesise that humans give causal (Lewis 1974) rather than descriptive explanations. Causal explanations give reasons for why there is a difference in plan quality, whereas descriptive ones describe what the differences are. Other work distinguishes between these types of explanations as answers to "why?" and "what?" questions (Miller 2019). We believe that causal explanations are more useful, and so expect to see these. H6 captures this hypothesis.

4.1 Methodology

In order to test our hypotheses, we designed an exploratory 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). Each participant was presented with four different planning problems. 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 model. However, we described the problem as if the participant were the modeller, so they had all of the environmental information that would be needed to model the planning problem. We also provided the participants with a visual description of the task. For each planning problem we presented users with the optimal plan to solve the problem. The plan was presented to the users in natural language and as a visual diagram. 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 worse quality. The participants were asked to complete the tasks described above for four distinct planning problems.

Delivery Task In the Delivery Task, a driver must deliver a package of meat to a butcher and cereals to a grocer. The initial state of the task and the travel time between locations (in minutes) are shown in Figure 1. It takes 10 minutes to drive the road between the depot and grocer and 15 minutes to drive the road between the grocer and butcher. It takes 20 minutes to drive across the bridge between the depot and butcher. It takes 1 minute to load packages into a truck, 1 minute for the driver to board the truck, and 2 minutes to unload and deliver packages to locations. 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 the bridge.

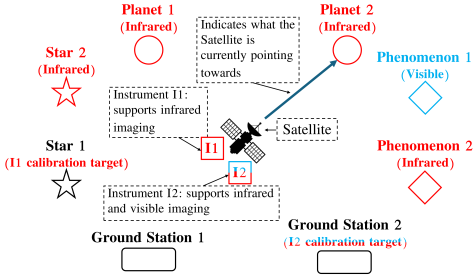
Figure 1: Initial state of the Delivery Task.



In this task, we asked the questions: "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 can be used instead of the unrefrigerated small truck T1 or refrigerated small truck T2?"

Satellite Task In the Satellite Task, there is a satellite that 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 2. 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 2: Initial state of the Satellite Task.

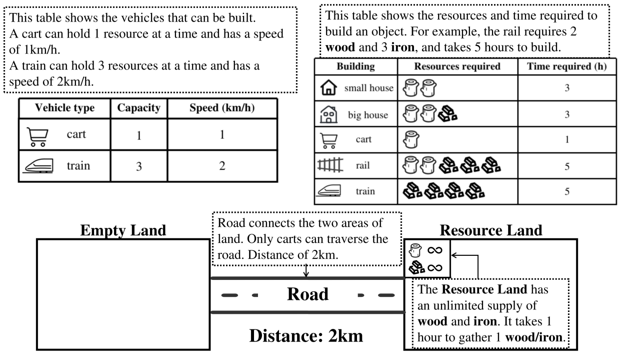


In this task, we asked: "Explain why it takes longer to solve the problem using instrument I1 to take the infrared image of Phenomenon 2 rather than instrument I2?" and "Explain why the problem becomes unsolvable when only instrument I1 can be used instead of instrument I2?"

Building Task In the Building Task, there are two locations 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, in Empty Land, small or big houses. Building a big house requires 1 more iron than a small house. A vehicle must be built to transfer resources between Resource Land and Empty Land. A vehicle can be either a cart or a train. 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 was presented to the study participants, shown in Figure 3. The goal is to build 2 houses of any type in Empty Land in the shortest possible time.

In this task, we asked the questions: "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 In the Rover Task, a rover must collect rock data from Location 1, collect soil data from Location 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 of the task is presented in Figure 4. The goal is to complete all 3 tasks in the shortest possible time.



This table shows the vehicles that can be built.
A cart can hold 1 resource at a time and has a speed of 1km/h.
A train can hold 3 resources at a time and has a speed of 2km/h.

Vehicle type	Capacity	Speed (km/h)
cart	1	1
train	3	2

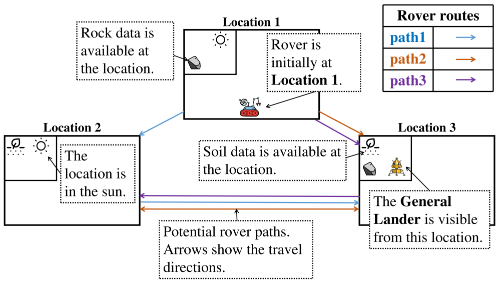
This table shows the resources and time required to build an object. For example, the rail requires 2 **wood** and 3 **iron**, and takes 5 hours to build.

Building	Resources required	Time required (h)
small house	🪵🪵	3
big house	🪵🪵🧱	3
cart	🪵	1
rail	🪵🪵🧱🧱🧱	5
train	🧱🧱🧱🧱	5

Empty Land
Road connects the two areas of land. Only carts can traverse the road. Distance of 2km.
Resource Land
∞
∞
Road
The **Resource Land** has an unlimited supply of **wood** and **iron**. It takes 1 hour to gather 1 **wood/iron**.
Distance: 2km

Figure 3: Initial State of the Building Task.

Figure 4: Initial state of the Rovers Task.



Rock data is available at the location.
Location 1
Rover is initially at **Location 1**.
Rover routes
path1
path2
path3
Location 2
The location is in the sun.
Soil data is available at the location.
Location 3
The **General Lander** is visible from this location.
Potential rover paths. Arrows show the travel directions.

In this task, we asked the following questions: "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 test our hypotheses that users tend to explain the difference in the quality of solutions to planning problems through the use of abstraction, 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 axial 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

We extracted the abstractions that can be used to form participants' explanations in order to test hypotheses H1 - H4. We did this through the use of the constructed planning models, using the Planning Domain Definition Language (PDDL) (Fox and Long 2003), and through the process of abstraction for extracting causes, described in Section 3.

For each explanation we identified the causal reason for the difference in the quality of the solutions that were presented to the participants. These reasons were properties of the task descriptions, we abstracted away the accompanying property in the PDDL model to determine if it was a cause and if it was complete or partial.

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 general 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, "Because 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 required for the train to operate. 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.". This is a descriptive explanation. The participant does not give a causal reason for why the rover must navigate more times using path 2 instead of path 1. 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 still 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: the presence of an extra navigation step in the plan.

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. Here, by source, we mean where the knowledge needed to produce the explanation was available. We identified nine different sources of knowledge used to produce these explanations and subsequently categorised the explanations by these nine sources through labelling. Six of these sources were explicit in that they were part of the descriptions of the problems given to participants. These sources were: problem description, abstraction information, the original, constrained and abstracted plan, and the question posed. Two of these sources were implicit in that they required reasoning to generate the information in the explanation, we considered contrastive and causal reasoning. The final source was extra information where the source was unclear.

We labelled a phrase in the explanation as from the original, constrained, or abstracted plan if it referred to actions that were present in those plans. The source of information of a phrase was categorised as from the original, constrained, or abstracted problem information 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. The source of information of 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 given 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 trucks 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 plan, the 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" and that is 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 involves 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 to be used, and why this makes the solution take longer.

6 Results and Analysis

From the eight questions for the four 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 whilst the shortest was 7 words. None of the explanations produced by participants had to be dismissed due to illegibility. Of these 160 explanations, 123 were causal explanations and 37 were descriptive explanations (H6). Performing a chi-square test (χ^2) with the null hypothesis (ϕ) related to H6, that the proportion of causal explanations and descriptive explanations is what you would expect by chance, $\chi^2(1, 160) = 46.23$, these results are significant at $p = 0.01$. We can reject ϕ and accept 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.

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 abstractions for 9 of the explanations because we could not classify the participants' answers as an explanation, or they were incorrect or difficult to understand. Of the explanations produced 94% corresponded to abstractions of the problem presented (H1). Performing χ^2 with ϕ related to H1, that the proportion of explanations that correspond to abstractions of the problem is equal to what you would expect by chance, $\chi^2(1, 160) = 126.03$, these results are significant at $p = 0.01$. We can reject ϕ and accept H1.

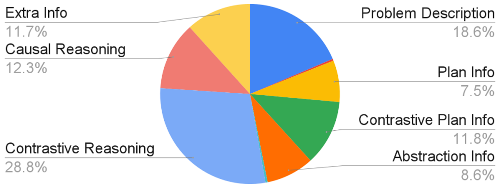
The abstractions extracted are shown in Table 1. The table shows the type of abstraction that was extracted, the abstraction, the code that was assigned through open coding, the effect of the abstraction, and the number of explanations that mentioned this abstraction. The abstraction and its type were extracted through axial coding of the codes in the table. The code was extracted from thematic analysis of the explanations. Then, through the use of the PDDL model 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 precondition abstraction 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 (22) 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 22 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. QC, quality complete, indicates that solutions to abstractions of both the original and constrained problems are equi-cost, and these are also equi-cost with the original (optimal) plan. C, complete, indicates that the abstraction caused both the original and constrained problems to produce equi-cost plans. We call both QC and C complete abstractions. P, partial, indicates that the abstraction reduced the difference in quality between the original and constrained problem solutions, which is a partial abstraction. I, irrelevant, indicates that the abstraction did not reduce the difference in the quality between the original and constrained problem solutions, we call this an irrelevant abstraction.

The majority of abstractions, 127, were QC abstractions. Only 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 ab-

Figure 5: Distribution of the sources of the explanations produced by participants.



stractions. 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 can be done in parallel with other actions so was not the reason for the difference in quality .

Of the 265 abstractions extracted 60% removed the difference in the quality of the solutions (H2). Performing χ^2 with ϕ related to H2, that the proportion of abstractions extracted removes the difference in the quality of the solutions to the original and constrained problems is what you would expect by chance, $\chi^2(1, 265) = 10.6$, these results are significant at $p = 0.01$. We can reject ϕ and accept H2.

Of the 265 abstractions extracted 48% cause the constrained problem to produce plans similar in quality to the original problem (H3). Performing χ^2 with ϕ related to H3, that the proportion of abstractions extracted causing the constrained problem to produce plans similar in quality to the original problem is what you would expect by chance, $\chi^2(3, 265) = 79.5$, these results are significant at $p = 0.01$. We can reject ϕ and accept 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). 71 explanations corresponded to one abstraction, 43 corresponded to two, 30 to three, 6 explanations corresponded to four abstractions, and 1 explanation corresponded to five abstractions. Performing χ^2 with ϕ related to H4, that the proportion of explanations corresponding to one, two, three, four, and five abstractions is what you would expect by chance, $\chi^2(4, 265) = 110$, these results are significant at $p = 0.01$. We can reject ϕ and accept H4.

6.2 Sources of Explanations

The sources of information for participant's explanations are shown in Figure 5. 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 implicit sources of information, explanations contained more information available through contrastive reasoning (28.8%) than causal reasoning (12.3%). From the explicit sources of information, explanations contained the most information from the problem description (18.6%). Explanations used little information from the abstracted plan (0.4%) and the question (0.3%).

It was not possible to assign a source for only 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, abstraction information, the original plan, the constrained plan, the abstracted plan, the question posed, contrastive and causal reasoning (H5). Performing χ^2 with ϕ related to H5, that the proportion of explanations formed from information available from these sources is equal to what you would expect by chance, $\chi^2(1, 23594, 160) = 13839$, these results are significant at $p = 0.01$. We can reject ϕ and accept H5.

7 Conclusions and Recommendations

Humans heavily use abstractions to explain the reason for the difference in the quality of plans (H1). These abstractions completely, rather than partially, explain the differences (H2) and caused the constrained problem to produce solutions similar in quality to the original solution (H3). The explanations that humans produce correspond to only a single abstraction (H4). Explanations contain information available in the task description and solutions they support (H5). This indicates that human-like-explanations can be generated from information from these sources. Humans produce causal rather than descriptive explanations (H6).

Although abstraction is a popular approach to explanation, there has been limited evidence that humans utilise abstraction to explain in an AI reasoning context. Support for H1 and H6 provide vital foundational support for the use of causal explanation via abstraction in this setting.

Prior work has not considered what abstractions produce the best explanations. We can recommend for human-like-explanations, from hypotheses H2 to H5, that abstractions should completely explain differences, cause the constrained problem to produce solutions similar in quality to the original solution, be minimal, and the explanation itself should contain information from the task description and solutions. In the future, we will use this information to create heuristics for finding abstractions that satisfy these properties and metrics for measuring their explanatory power, which we will independently evaluate.

Table 1: Abstractions extracted from participants' explanations for questions 1 and 2 of the Satellite, Rovers and Building Tasks.

Type	Abstraction	Code	Effect	Count
Satellite Task Question 1				
Precondition	Can take image	I1 does not support visible mode	QC	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	QC	2
Constraint	Use I1	I2 should take all images, instead of I1	QC	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	QC	18
Building Task Question 1				
Precondition	Connected By Rail	A rail is needed to use the train	C	5
	Is Train	You have to build the train, which takes more resources and time	C	5
Duration	Find Resource	More time spent gathering resources	QC	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 And Build Rail	Requires more resources to build the train and the rail	QC	2
	Available Iron	It requires more iron to build the train and rail	QC	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/moves more resources, 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	QC	2
	Available Resources From Build Big House	More resources required to build big houses	QC	6
	Available Iron	More iron required	QC	1
	Available Iron From Build Big House	More iron required to build big houses	QC	7
Rovers Task Question 1				
Precondition	Lander Visible	Can only communicate to the general lander from location 3	QC	4
	In Sun	There is no sun in location 3/can not reacharge	I	2
Duration	Navigate	Takes more time to navigate	QC	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 general lander from location 3	QC	13
	In Sun	There is no sun in location 3/can not recharge	I	8
	Can Traverse	Can not go back to location 3	QC	1
Function	Energy	Energy level	I	3

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.

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.; and Long, D. 2003. PDDL2.1: An Extension to PDDL for Expressing Temporal Planning Domains. *Journal of Artificial Intelliigence Research*, 20: 61–124.

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.; 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.